Chapter 13

Statistics

Statistics

PREFACE

For appropriate process and quality control in the cement and concrete industry, a large number of data are derived. Optimum benefit is, however, only achieved if these are adequately processed and interpreted. Statistics is one of the important means to make best use of the data be it by application of numerical methods and/or by graphical representation.

The present handbook describes the relevant basic definitions, formulae and applications of statistical methods which are useful in the cement industry. Emphasis is put on adequate data description and graphical representation to ensure reasonable processing and interpretation of statistical data. Most of the described procedures are illustrated with practical examples.

Chapters 1 and 2 are concerned with the basic ideas of statistics, the rules for the representation of a given set of data (graphical representation, numerical measures) and the treatment of outliers.

Chapters 3 to 8 present some statistical methods, useful for decision making, experimentation and process control.

Of special practical significance is the Application Section (chapter 5.2), which includes a procedure manual with a general check list and a collection of important test procedures. Chapter 9 gives an outwork to more sophisticated statistics.

Appendix I contains a selection of practical examples to illustrate applicability and interpretation of the demonstrated methods. A useful work sheet to construct frequency tables and to check the data for normality is given in Appendix II. Further Appendices contain: the required statistical tables for the determination of confidence limits and the application of test procedures, a list of recommended literature and an index of examples used in the text.

A subject index (English, German, French) of statistical terms is provided in Appendix VI.

1. INTRODUCTION

Statistics is concerned with methods for collecting, organising, summarising, presenting and analysing data, obtained by measurement, counting or enumeration.

With descriptive statistics a given set of observations is summarised or presented to get a quick survey of the corresponding phenomena.

In a more sophisticated analysis of representative samples statistical inference (inductive statistics) allows conclusions to be drawn about the entire population.

A sample is considered to be representative only if it is drawn from the population at random. Such a group of n observations is called a sample of size n.

Some main topics of interest in statistical analysis are:

a) Location of the data:
Where are the observations located on the numerical scale? This question leads to the use of central values as the mean or the median.

b) Variability or dispersion:
Problems concerning the degree to which data tend to spread about an average value.

c) Correlation:
Degree of dependence between paired measurements, e.g.: Is there a real dependence of mortar strength on the alkali content in clinker?

d) Regression:
Fitting lines and curves to express the relationship between variables in a mathematical form especially used for prediction and calibration.

e) Splitting variability:
Looking for the importance of several causes to the variability of observations. Variability arises due to different components of random errors. Special experiments must be designed to split these components.

The use of statistical methods and the interpretation of statistical results requires a certain comprehension of variability. If ten pieces of coal from a delivery are analysed as to their water content, the results are not identical with the water content of the full quantity. Rather we have ten different results in a certain range. This variability of data arises not only from the fact that the pieces really have different water contents, but also because the results are influenced by three different types of errors which may occur in every set of observations:

Random errors cannot be avoided. They are due to imprecise measurement, rounding of data, environmental effects and not identically repeatable preparation. The amount of random errors may be expressed by statistical measures of variability.

Systematic errors lead to a bias in location, but not necessarily in dispersion. A bias in dispersion is obtained if several systematic errors are mixed.

Example: Every laboratory assistant has usually his own systematic error. This error may be relevant or not and it may change with time. But we always expect a greater variability of the observations if more than one person have performed the measurements. On the other hand, we recognise in this example that a small variability does not necessarily lead to a better average value. Environmental effects may be systematic too (e.g. air humidity).

Gross errors are "wrong" values in the set of observations. Experience shows that 5% to 10% of gross errors have to be expected in a data set. Reasons may be: wrong reading of scales, errors in copying data, data not legible, miscalculations, gross error in measurement. Gross errors have a considerable influence on statistical results.

Careful measurement and data handling is, therefore, important. Data have to be inspected for outliers before a statistical analysis is performed.

Conclusion: The deviation of the single results from the true mean of the full quantity originate from real differences between the samples on one hand, and from the occurrence of several types of errors on the other.

It is often stated that "everything can be proved with statistics", this reasoning clearly is wrong; false or often only misunderstood statistics originate from insufficient representations, application of wrong procedures or assumptions, or misleading interpretation of the results. Especially graphical representations can easily be manipulated. It is therefore essential that the applicant of statistical techniques knows what he can and what he cannot do! An amusing booklet dealing with such statistical "lies" is "How to Tell the Liars from the Statisticians".

In our days of growing computer use, a lot of powerful (and sometimes less powerful or even poor) statistical software packages are available, leading to extensive use of statistics procedures by nonstatisticians or people not having enough statistical background. These packages manage nearly any instruction without being able to decide whether the statistical procedure is appropriate. This bears a great danger of use of statistics by statistical amateurs or ignorants. It is therefore absolutely necessary that the user of statistics has a solid statistical education.

2. SIMPLE DATA DESCRIPTION

In this section some descriptive methods are presented to get a quick survey of the data. The methods are illustrated with an example of concrete strength.

Example 1

The following data represent the compressive strength of 90 cubes (20 x 20 x 20 cm) of concrete. The data listed in chronological order of measurement is called set of observations.

35.8	39.2	36.8	32.4	30.7	30.8	23.5	22.8	23.7	31.7	34.6	27.6	29.9	28.4	29.3
33.0	37.6	38.1	33.3	38.9	37.1	33.3	33.4	36.4	44.3	48.9	40.1	43.4	35.4	36.6
32.8	34.1	37.4	27.9	30.2	32.0	45.3	45.8	41.0	26.1	27.9	24.4	35.3	34.5	36.1
30.1	40.2	37.9	25.0	23.0	27.8	33.5	34.2	30.0	29.0	35.2	35.8	23.9	34.9	31.5
35.9	39.7	39.4	32.4	33.6	35.2	32.8	30.2	31.6	28.5	28.5	30.3	31.4	31.8	35.5
27.1	24.5	20.9	24.6	27.2	31.7	32.2	38.6	32.8	37.8	36.8	35.3	41.9	34.4	35.5

Generally a set of n observations is denoted by $x_1, x_2, ..., x_n$ where the index j of x_j corresponds to the number of the observation in the set.

2.1 Graphical Representation

Although the set of observations gives the complete information about the measurements, it is little informative for the reader. A better survey is obtained by grouping the data in classes of equal length, presented in a frequency table with tally.

Rules for classification:

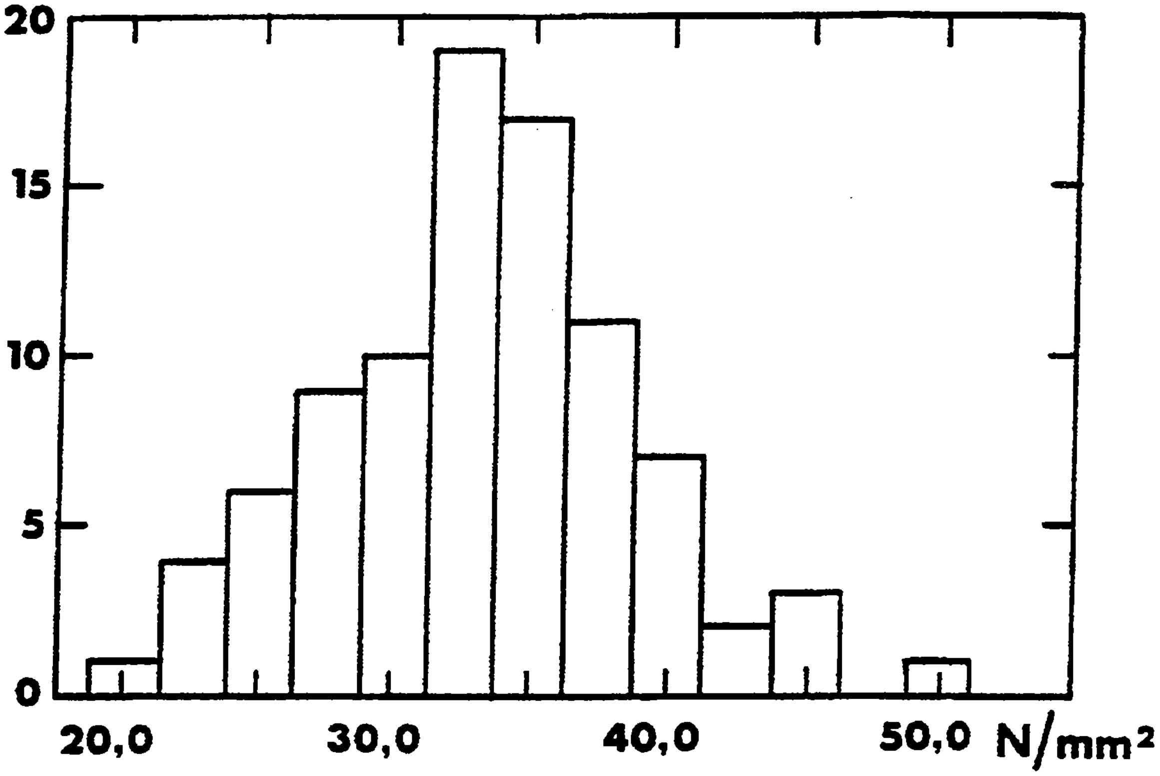
1) The mid points should be impressive values with a few number of digits.
2) Choose 8 to 20 classes of equal length (approx. $\sqrt{n}$; classes)
3) Class boundaries should not coincide with observed values (if possible)

mid-point	class	tally	absolute frequency	relative frequency
20.0	18.75 - 21.25	I	1	0.011
22.5	21.25 - 23.75	IIII	4	0.044
25.0	23.75 - 26.25	卌 I	6	0.067
27.5	26.25 - 28.75	卌 IIII	9	0.100
30.0	28.75 - 31.25	卌 卌	10	0.111
32.5	31.25 - 33.75	卌 卌 卌 IIII	19	0.212
35.0	33.75 - 36.25	卌 卌 卌 II	17	0.189
37.5	36.25 - 38.75	卌 卌 II	11	0.122
40.0	38.75 - 41.25	卌 II	7	0.078
42.5	41.25 - 43.75	IIII	2	0.022
45.0	43.75 - 46.25	II	3	0.033
47.5	46.25 - 48.75		0	0.000
50.0	48.75 - 51.25	I	1	0.011
Total			90	1.000

In this representation we recognise directly a minimum of 20.0 N/mm^2, a maximum of 50.0 N/mm^2, and an average value between 32.5 and 35.0. The measurements are distributed symmetrically about the average value.

The graphical representation of the tally with rectangles is called a histogram.

Fig. 1 **Histogram.**



$$\bar{x} = 33{,}2\ N/mm^2$$

$$\tilde{x} = 33{,}5\ N/mm^2$$

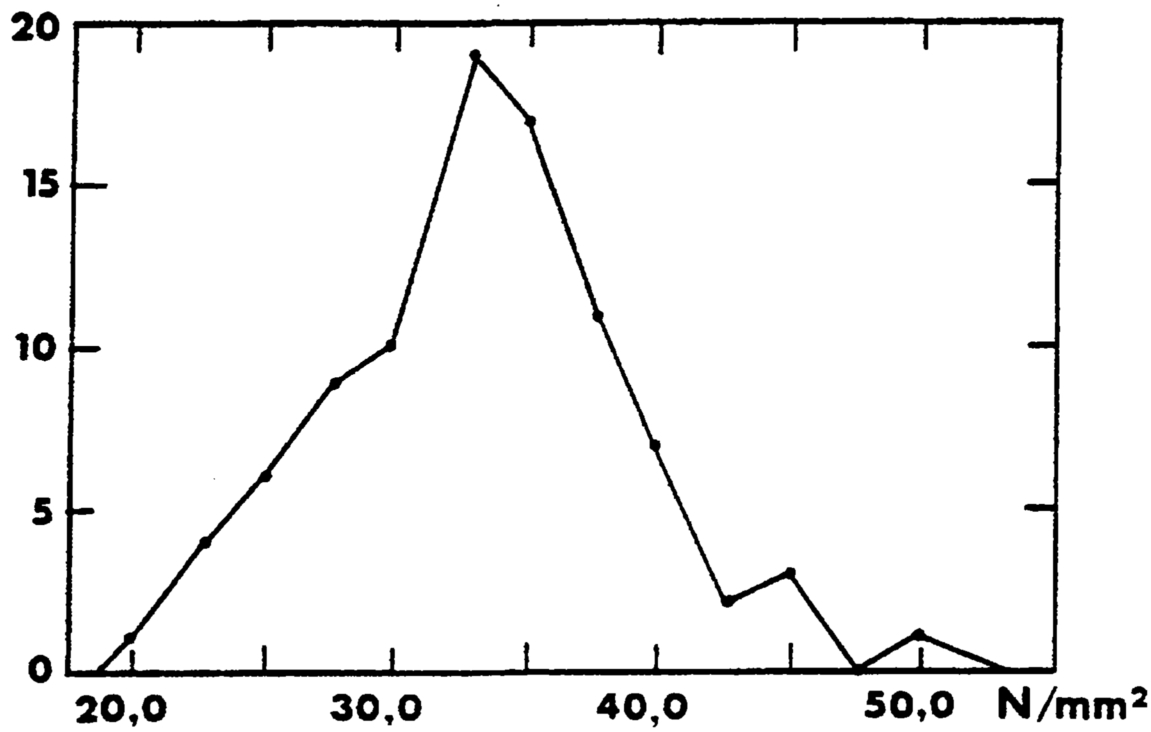
The area of the rectangles must be proportional to the tally not the height of the bar, e.g. if two classes contain the same number of observations and the width of class 1 is twice the one of class 2, then the height in class 1 is half the height in class 2. Same number of observations same area of rectangle!

Two further informative graphs are the frequency curve and the cumulative frequency curve. They are constructed as follows:

Frequency curve

- plot the class frequency (absolute or relative) against the class midpoint (again take into account the note for the histogram, same number = same area).

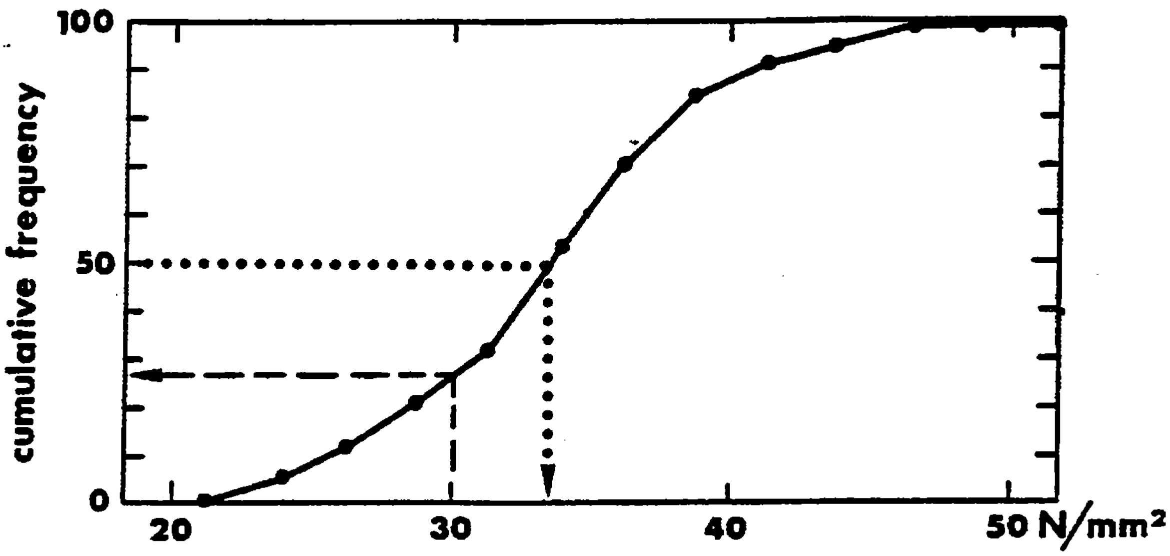
Fig. 2 **Frequency Curve.**



Cumulative frequency curve

- cumulate the absolute frequencies less than the upper class boundary for every class
- plot the cumulative frequency in percent against the upper class boundary.

Fig. 3 Cumulative Frequency Curve.

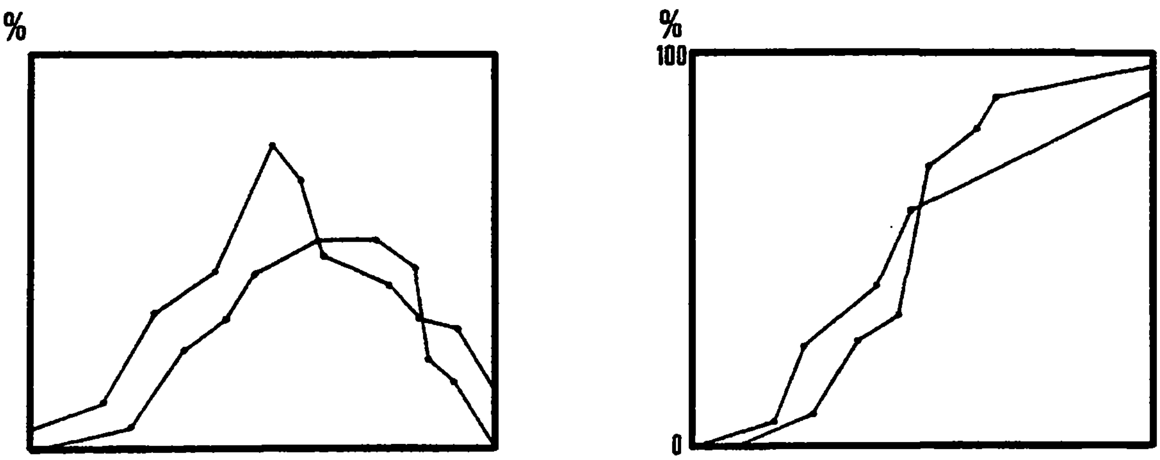


From this graph we can determine the portion of observations smaller than any given strength value x.

Example: Percentage of the observations smaller than 30.0 N/mm^2: 26 %. Inverse problem: Half of the measurements are smaller (resp. greater) than 33.5 N/mm^2.

These two frequency curves are especially suited to compare two or more different distributions with one another.

Fig. 4 Compare Distributions.



Stem and Leaf Plot

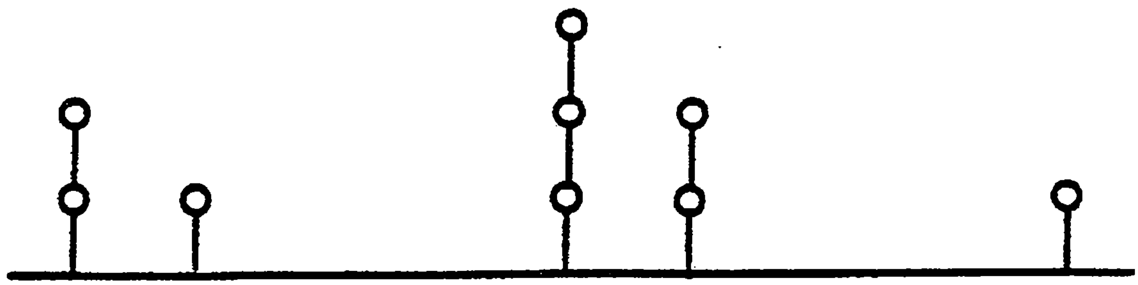
The stem and leaf plot is very similar to the histogram but allows to reconstruct the individual data. [The stem and leaf plot is shown in the next picture, the comment does not belong to it, but shows how the data are reconstructed.]

STEM	LEAF	COMMENT
20	9	20.9
21		no observation
22	8	22.8
23	0579	23.0, 23.5, 23.7, 23.9
24	456	24.4, 24.5, 24.6
25	0	
26	1	etc.
27	126899	
28	455	
29	039	
30	0122378	
31	456778	
32	0244888	
33	033456	
34	124569	
35	2233455889	
36	14688	
37	14689	
38	169	
39	247	
40	1	
41	029	
42		
43	4	
44	3	
45	38	
46		
47	9	
48		
49		
50		

Small Samples

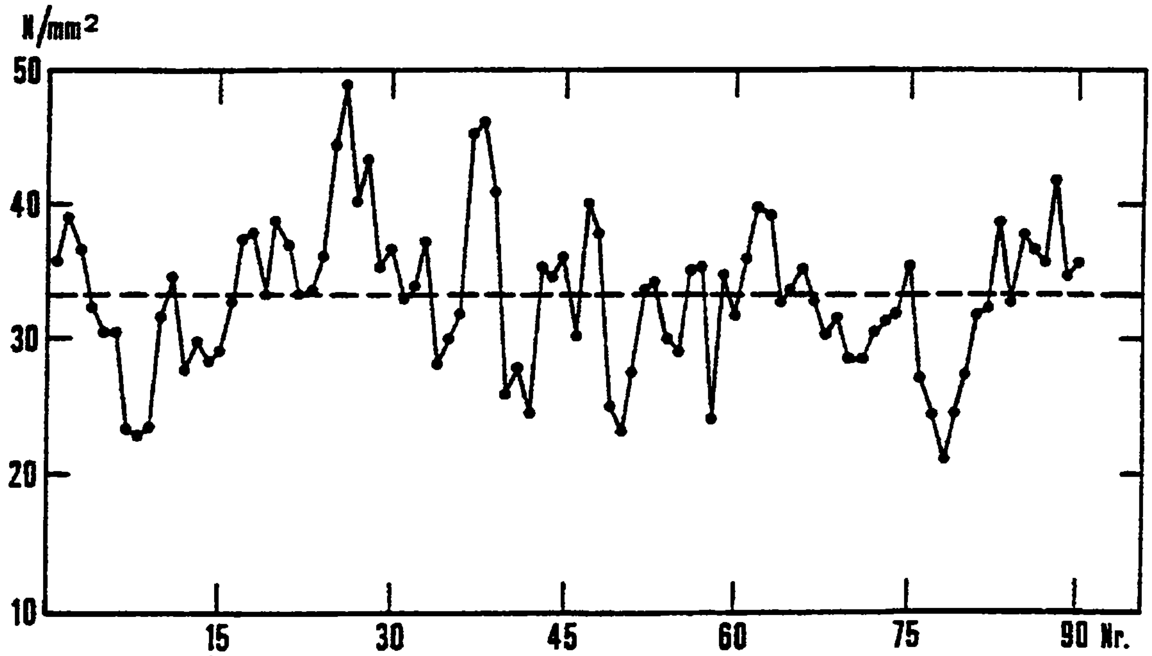
Individual values are marked on the scale

Fig. 5 Small Sample Table.



A further attractive possibility to describe graphically the distribution of a variable, the Box Plot, will be given in the next section. Often it is appropriate (especially in quality control) to plot the observations in chronological order to show a possible change of the level during the experiment.

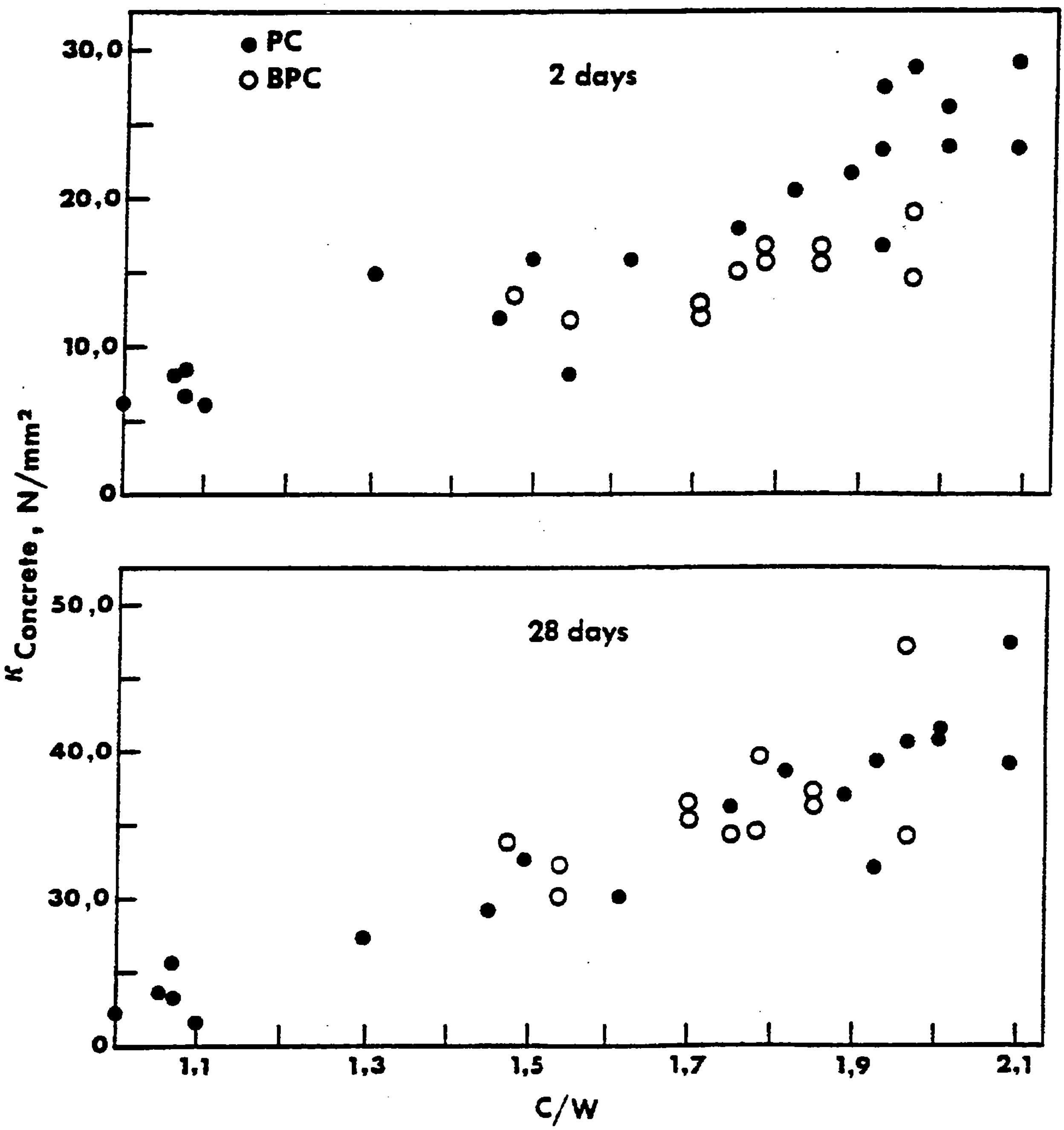
Fig. 6 Chronological Order.



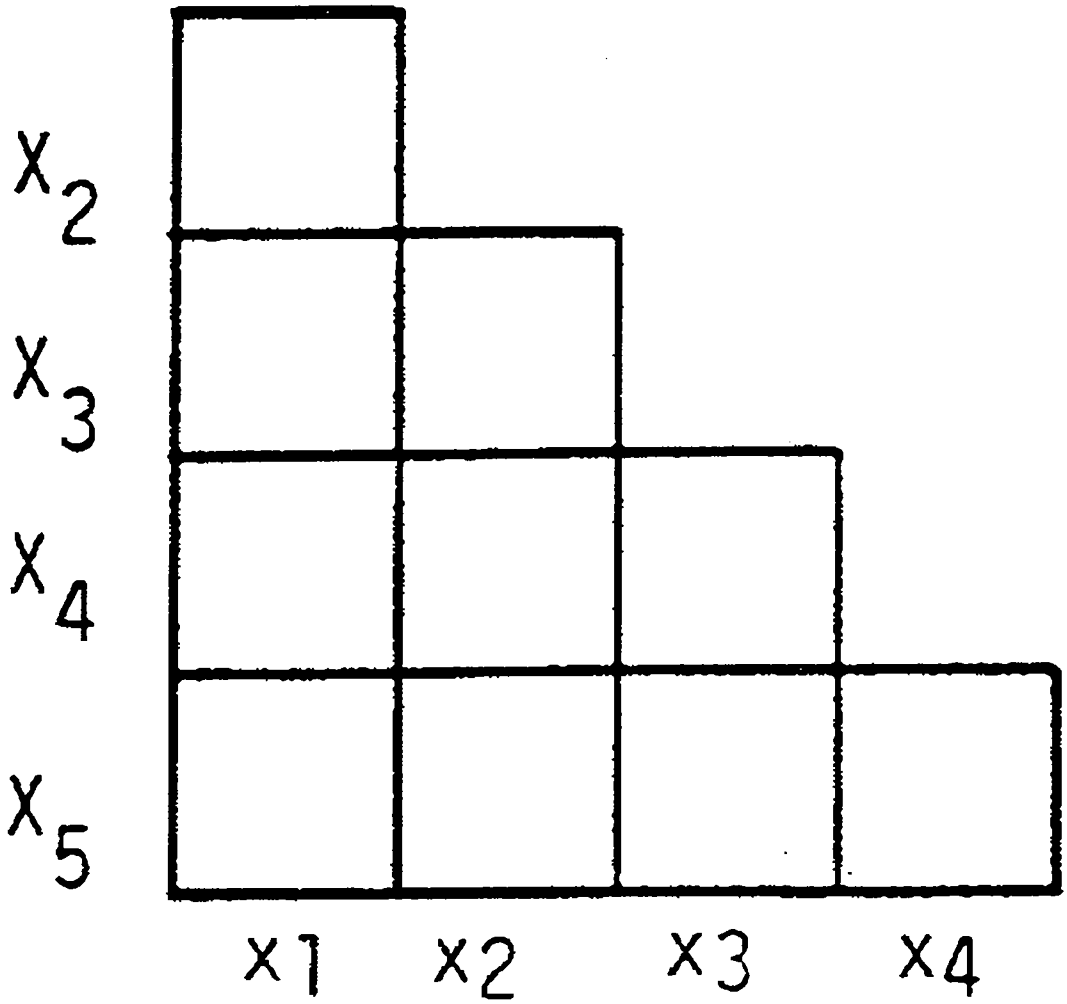
If only few data are available, the single values may be represented as points on the measurement scale (cf. example A1, Appendix I).

The scatter diagram or scatter plot is used to represent the relationship between two variables (paired observations concerning the same individual sample). The following diagrams show the dependence of concrete strength in N/mm^2 from cement/water ratio after 2 days and 28 days. Different symbols can be used to discriminate groups of observations. In the present case two groups are considered: Portland cement (PC) and blended Portland cement (BPC).

Fig. 7 Scatter Diagram.



If we are dealing with more than 2 variables a suggested graphical representation is the draftmansplot, which plots all pairwise scattergrams in one picture.


X2
X3
X4
X5
X1
X2
X3
X4

2.2 Statistical measures

As statistics deals with the behaviour of random variables, i.e. variables which take certain values with certain probabilities, we are interested in describing this behaviour of the random variable by a few characteristic measures. The first measures we will be considering are the p quantiles (or p fractiles), which are roughly said those values for which 100.p % $(0 <= p <= 1)$ of the data are smaller or equal to. The p quantiles allow a full description of the data, similar e.g. to the cumulative frequency curve from Section 2.1. The use of the p quantiles gives way to construct a simple and attractive graphical representation of the data, the box plot.

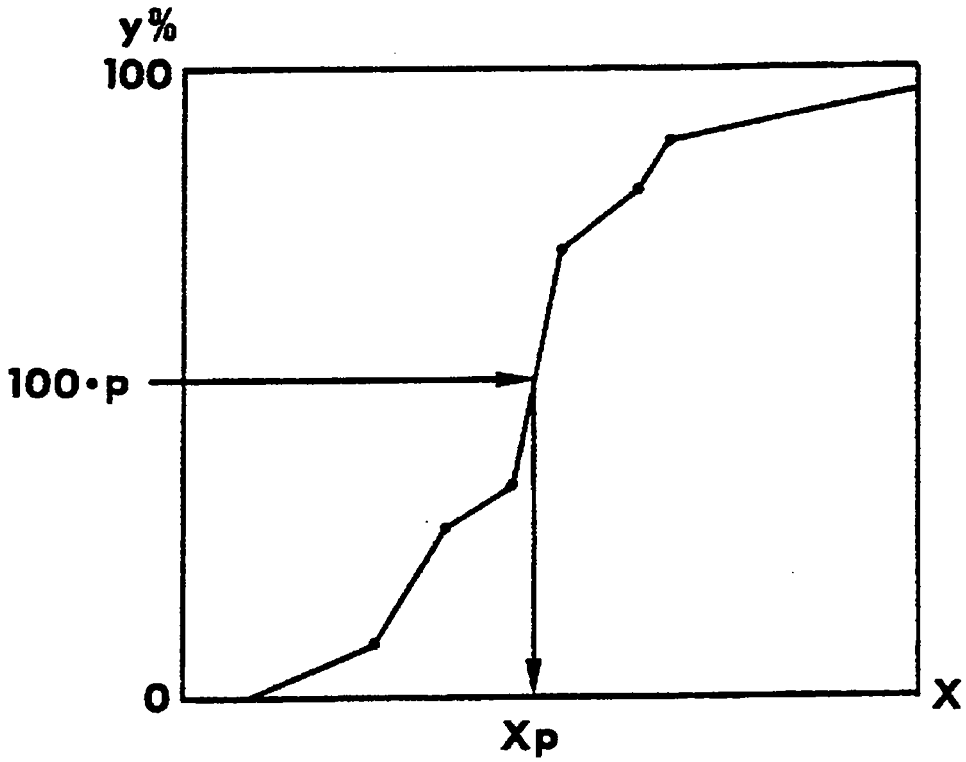
2.2.1 The p-quantile

x_p is called the p-quantile (or p-fractile) if

$100 \bullet p\%(0 <= p <= 1)$

of the values of the random variable are smaller or equal to x_p.

Based on a set of observations from a random variable x, a first method to estimate these quantiles is given by the cumulative frequency curve.



Given a sample x(i), i=1,..,n the empirical cumulative distribution function (cdf) is defined as

F(x): = (number of the x(i)'s smaller or equal to x) / n

and the following relations between the p quantile y(p) and the cdf F(x) hold

a) F(y(p)) = p

b) Y(p) = min {x(i) : F(x(i)) >= p}

The p quantile y(p) is roughly said the value for which 100*p% of the observed data are smaller or equal to.

Estimation of P quantile v(P)

Denote by z(i), i=l, ..,n the ordered sequence of the x(i)'s, e.g.

z(1) <= z(2) <= z(3) <= .... '= z(n 1) '= z(n)

and let p(i) = (i 0.5)/n.

Then the P quantile can be computed as follows:

1) If P equals a p(i), then z(i) is the P quantile

Y(P) = z(i)

2) Otherwise compute

j(P) = nP + 0.5

and split this value j(P) into a whole number j and the remaining part B (e.g. 1.347 is split into j=1 and B=0.347). The P quantile is then

$$y(P)=(1-B)z(j)+Bz(j+1)$$

An example

Consider the following ordered sample

i	Z(I)	p(i)
1	156	0.0417
2	158	0.1250
3	159	0.2083
4	160	0.2917
5	161	0.3750
6	161	0.4583
7	163	0.5417
8	166	0.6250
9	166	0.7083
10	168	0.7917
11	172	0.8750
12	174	0.9583

12.5 % quantile (P=0.125)

As P=p(2), our 12.5 %-quantile y(.125) equals the second ordered observation, thus y(.125) = 158

10 % quantile (P=O.1)

Compute j(P)=nP + 0.5 = 12*0.1 + 0.5 = 1.7

As this yields no whole number our 10 % quantile is (j=1 and B=0.7)

y(.1) = (1 B) z(1) + B z(2)

= 0.3*156 + 0.7*158 = 157.4

25 %-quantile (P=0.25)

Compute j(P)=nP + 0.5 = 12*0.25 + 0.5 = 3.5

As this yields no whole number our 10 % quantile is (j=3 and B=0.5)

$$y(.25) = 0.5*159 + 0.5*160 = 159.5$$

50 % quantile (P=0.5)

Compute j(P)=nP + 0.5 = 12*0.5 + 0.5 = 6.5

As this yields no whole number our 10 % quantile is (j=6 and B=0.5)

$$y(.5) = 0.5*161 + 0.5*163 = 162$$

Some of the quantiles have special names, e.g.

50 % quantile	median (M)
25 % quantile	lower quantile (LQ)
75 % quantile	upper quantile (UP)
10 % quantile	lower decile (LD)
90 % quantile	upper decile (UD)
OQ UQ	Inter quantile range (IQ)

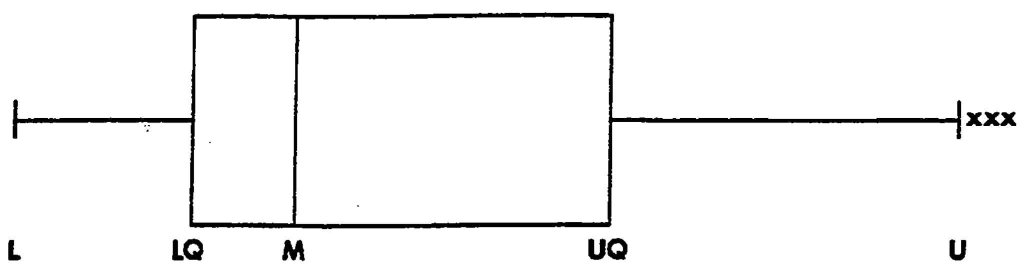
2.2.2 The Box Plot

Today the box-plot is perhaps the most frequently used graphical representation for univariate data. It offers a condensed picture of the data's distribution and shows location, variability and extremes of the data.

In order to construct the box plot we need three quartiles:

- the median
- the lower quantile
- the upper quantile

The construction is very simple:



where L = max (LQ 1.5 x IQ, min x_i)

U = min (UQ + 1.5 x IQ, max x_i)

IQ = UQ IQ

x denotes extreme values.

This construction is based on the fact that for a normal distribution approximately 1 % of the data lie outside the interval (L,U).

An alternative method also frequently used, would be not to use L and U but instead use the lower and upper decile. Especially for non normal distributions this representation seems preferable.

2.2.3 Measures of location

The arithmetic mean is the mean of all observations and corresponds to the centre of gravity in physics.

$$\bar{x} = \frac{x_1 + x_2 + \dots x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{x} = \frac{35.8 + 39.2 + 36.8 + \dots + 34.4 + 35.5}{90} = 33.199 N/mm^2$$

The median is the central value. Half of the observations are smaller, respectively greater than the median.

Place the measured values in an ordered array starting with the smallest value

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

From this array we obtain the median by

$$\tilde{x} = x_{\frac{(n+1)}{2}} \qquad \text{if n is odd}$$

$$\tilde{x} = \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)} \right) \qquad \text{if n is even}$$

For a great number of observations the ordering is very laborious. In this case the median may be determined graphically from the cumulative frequency curve by looking for the point on the x axis corresponding to a cumulative frequency of 50 %.

Mean or Median?

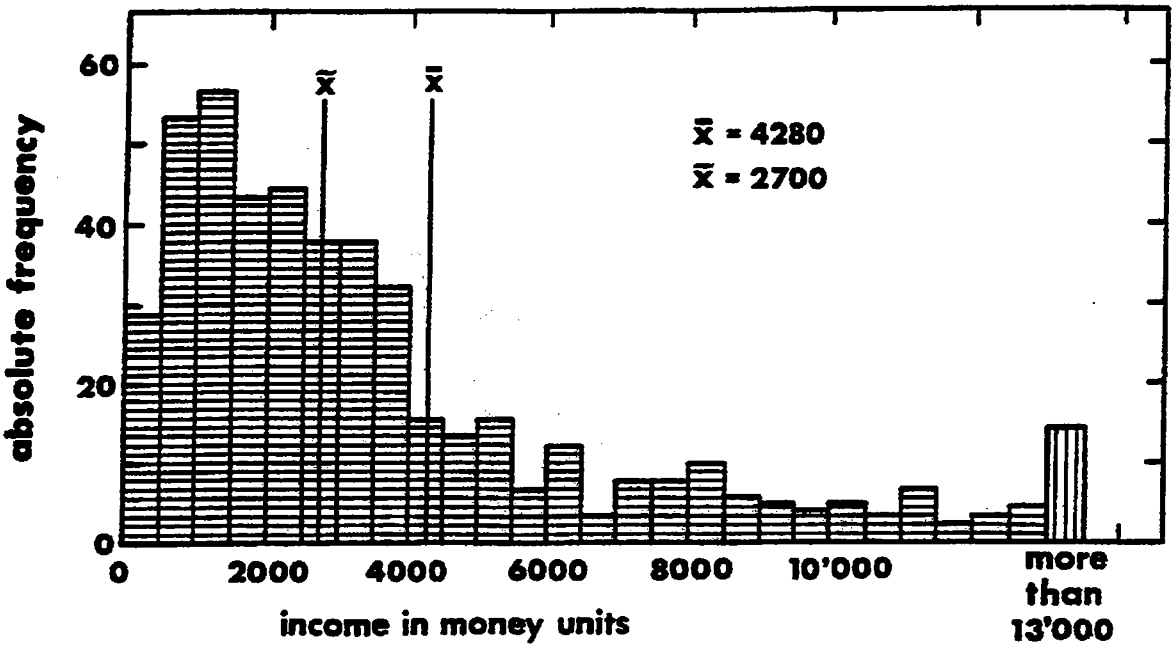
If the histogram is symmetrical, mean and median are approximately the same. In this case the mean is a better estimate of central tendency if the distribution is not too long tailed.

If the histogram is skew, mean and median are different.

- Use the mean if you are interested in the centre of gravity or the sum of all observations
- Use the median if you are interested in the centre, i.e. with equal probability a future observation will be smaller, resp. greater than the median.

Example 2

The following histogram represents the income of 479 persons



In this extremely skew distribution the arithmetic mean is quite different from the median. Whereas the median represents a typical income, the mean can be used in a projection to estimate the total income of the population if the sample is drawn at random.

The trimmed mean is used to estimate the mean of a symmetrical distribution if gross errors in the data are suspected.

To calculate the α-trimmed mean the α-percent largest and smallest values are deleted. The trimmed mean is then the arithmetic mean of the remaining observations. Notation for the 5%-trimmed mean: $x_{\text{trimmed 0 05}}$

Usual values for α: 5 % to 10 %

The weighted mean is used if certain weighting factors w_i are associated with the observation x_i. Reasons may be: - samples of unequal weights - observation are measured with unequal precision

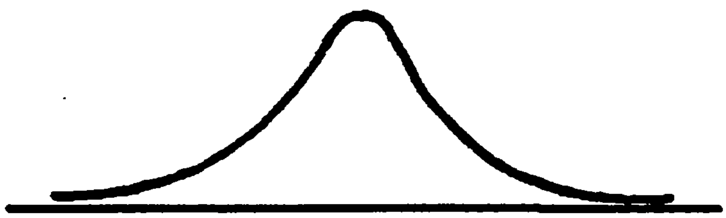
$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$$

Schematic survey on the use of measures of location:

Type of distribution

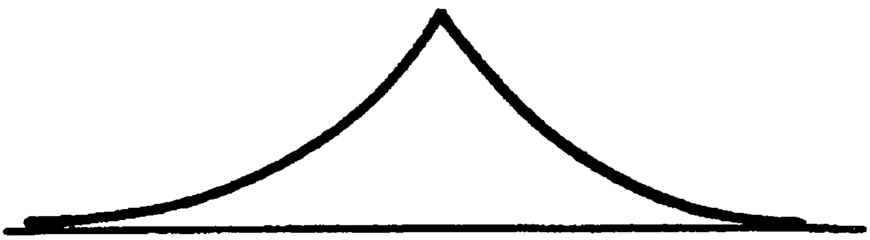
normal   shorttailed

Measure to be used: arithmetic mean $\bar{x}$



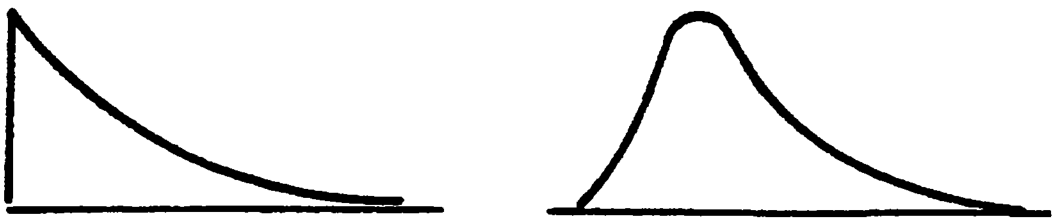
longtailed, symmetrical

Measure to be used: trimmed mean $\bar{x}_{trimmed}$



extremly longtailed

Measure to be used: median $\tilde{x}$



skew distribution

Measure to be used: median or arithmetic mean. The choice depends on the interest of the user.

Remark: The median is equal to the 50%-trimmed mean.

2.2.4 Measures of Variability

The variance is the mean square deviation of single observations from the mean.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

For practical computations use

$$s^2 = \frac{1}{n-1}\left(\sum_i x^2 - \frac{1}{n}(\sum x_i)^2\right)$$

The standard deviation is the positive square root of the variance.

$$s = +\sqrt{s^2}$$

In our example:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 55.596 N/mm^2$$

The coefficient of variation is a relative measure of the variability (relative to the mean).

$$v = \frac{s}{\bar{x}}$$

v is often used for comparing variabilities. It is especially useful if s increases proportionally with the mean $\bar{x}$, i.e. v = constant.

Example of compressive strength:

$$v = \frac{5.596}{33.199} = 0.169 = 16.9\%$$

The range is the difference between the largest and the smallest value.

$$R = X_{(n)} - x_{(1)} = X_{max} - X_{min}$$

Useful with small sample sizes.

Note: The range increases in general with increasing sample size n (greater probability to get extreme values).

If the variables under consideration are assumed to be normally distributed then

the range can be used to provide an estimate of the standard deviation, according to

$s = R/k$ where $k \cong \sqrt{n}$ for $3 \leq n \leq 10$

For sample size n >10 divide the set of observations in random sub samples of size n', calculate the arithmetic mean $\overline{R}$ of the range in these sub-samples, and estimate s according to

$$s \sim \overline{R}/k$$

using a k-value corresponding to n'. . For a first rough estimation the following k values may be used:

$k \sim 4$	for n = 20
$k \sim 4.5$	for n = 50
$k \sim 5$	for n = 100

The interquartile range is the difference between the upper and lower quartiles.

$$Q = x_{.75} - x_{.25}$$

If the distribution is symmetric $s \approx Q.\frac{4}{3}$

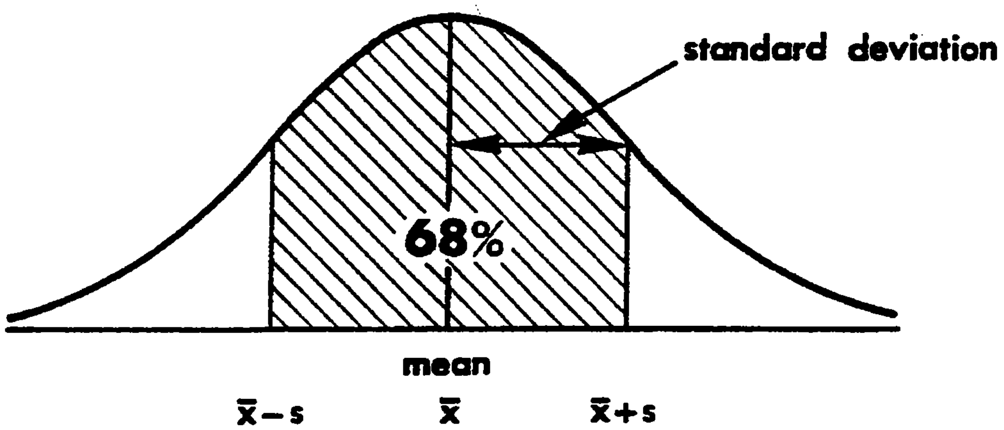
2.2.5 Statistical program packages

There are a lot of statistical programs on PCs, microcomputers and mainframes that offer the possibility to compute these statistical measures including quartiles and even to plot the box plots. The perhaps most prominent under many others are SAS, SYSTAT/SYGRAPH, SPSS, BMDP, STATGRAPHICS.

2.2.6 Interpretation of the standard deviation

In the histogram of example 1 (chapter 2.1) frequency of observations decreases on both sides of the mean. The distribution law seems to be symmetrical. In the present case we consider the 90 measurements to be a random sample of a specific distribution model, the normal distribution (or Gaussian distribution). This type of distribution is represented by a symmetrical, bell shaped curve, and often observed in practical applications.

The normal distribution is determined by the mean and the standard deviation.



For this distribution about 68 % of the observation are expected in the interval mean ±1 standard deviation.

If the distribution is not normal, the standard deviation allows no direct interpretation. In this case we can determine the interval that contains a certain portion of the observations graphically from the cumulative frequency curve.

Another possibility to get an interpretation of the standard deviation in a skew distribution is "normalising" of the distribution by transformations. In the case of example 2 (chapter 2.2.3), the logarithms of incomes show approximately a normal distribution. So called variance-stabilising transformations are especially used to fulfil normality conditions in higher statistical analysis.

For further information consult e.g. Natrella (1963).

Further properties of the normal distribution are given in section 3.

2.2.7 Outliers

As mentioned in the introduction we have to expect about 5 % to 10 % gross errors in a set of observations. Most of them may not be recognisable in the region of all other values. Some may be extremely outlying values with an important influence on statistical results.

In modern statistics robust methods are studied, which are not sensitive to a certain portion of gross errors, as for example the median or the trimmed mean. More sophisticated robust procedures are in general rather complicated.

If classical measures as the standard deviation and the arithmetic mean are used, we have to check the data for outliers and to eliminate them from the set of observations (cf. example A2, Appendix I).

Note: If outliers are eliminated, they must be recorded separately in the report.

To detect outliers, check the data plot (tally, histogram or chronological order) for suspicious values. If outliers are suspected and the data show a normal distribution, use the Dixon criterion for rejecting observations ($n<26$).

The Dixon Criterion

Procedure

1) Choose a, the probability or risk we are willing to take of rejecting an observation that really belongs in the group.

2) If

$3 \le n \le 7$	Compute r_{10}
$8 \le n \le 10$	Compute r_{11}
$11 \le n \le 13$	Compute r_{21}
$14 \le n \le 25$	Compute r_{22}

where r_{ij} is computed as follows

r_{ij}	If $X_{(n)}$ is suspect	If $X_{(1)}$ is suspect
r_{10}	$(X_{(n)} - X_{(n-1)})/(X_{(n)} - X_{(1)})$	$(X_{(2)} - X_{(1)})/(X_{(n)} - X_{(1)})$
r_{11}	$(X_{(n)} - X_{(n-1)})/(X_{(n)} - X_{(2)})$	$(X_{(2)} - X_{(1)})/(X_{(n-1)} - X_{(1)})$
r_{21}	$(X_{(n)} - X_{(n-2)})/(X_{(n)} - X_{(2)})$	$(X_{(3)} - X_{(1)})/(X_{(n-1)} - X_{(1)})$
r_{22}	$(X_{(n)} - X_{(n-2)})/(X_{(n)} - X_{(3)})$	$(X_{(3)} - X_{(1)})/(X_{(n-2)} - X_{(1)})$

3) Look up $r_{1-\alpha/2}$ for the r_{ij} from Step (2), in Table A 2

4) If $r_{ij} > r_{i-\alpha/2}$ reject the suspect observation; otherwise, retain it.

In the case of a sample size n>25 use the following procedure:

1) Choose α, the probability or risk we are willing to take of rejecting an observation that really belongs to the group

2) Calculate $Z_B = \dfrac{x_{(n)} - x_{(1)}}{s}$

3) Look for $Z_{T,1-\alpha}$ in Table A 3, Appendix III

(4) If $Z_b > Z_{T,1-\alpha}$ reject the suspect observation, otherwise retain it.

Note: The presented outlier rejecting rules are only valid in normal distributions. In a skew distribution, the elimination of outliers is very dangerous and should be avoided. In this case the reason for the extreme observation must be known.

A check for outliers is not necessary if the trimmed mean or the median is used and if we are only interested in a location measure.

For further tests on outliers cf. "Wissenschaftliche Tabellen Geigy, Statistik".

The standard deviation is extremely sensitive to outliers.

A check is therefore important, because it is not allowed to calculate a standard deviation from a trimmed set of observations.

3. THE NORMAL DISTRIBUTION (ND)

As mentioned in section 2, the normal distribution is a theoretical model of a statistical universe (population), defined by the mean μ and the standard deviation σ. Graphically it is represented by a smooth, symmetric mean x and the standard deviation s of the samples are estimated for the true, but unknown values μ and σ.

By the standardisation formula

$$Z = \frac{x-\mu}{\sigma} \text{ (respectively } \frac{x-x}{s} \text{ if n is large)}$$

the ND is transformed in a normalised form with mean 0 and standard deviation 1.

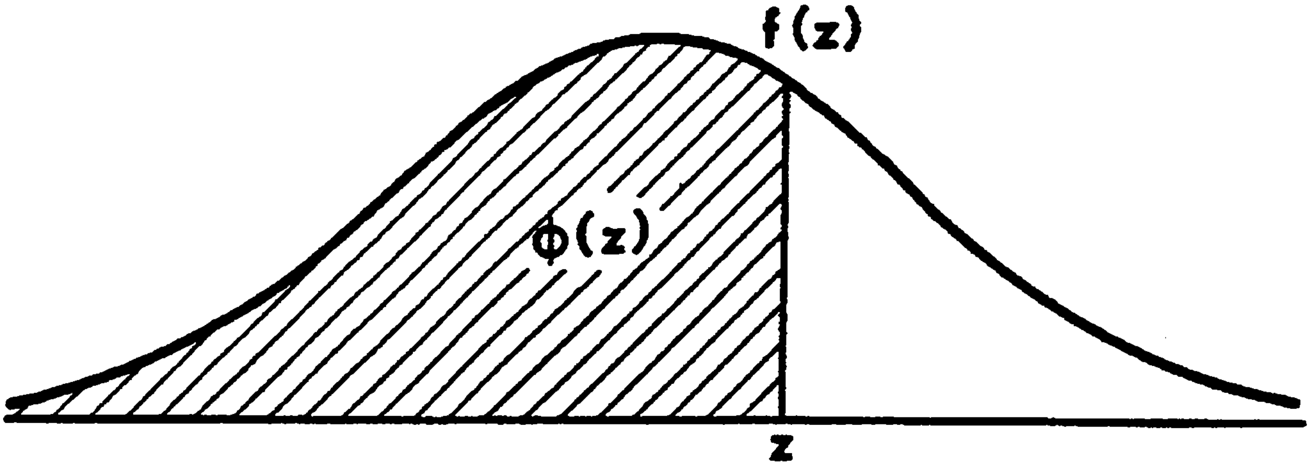
Formulae for the standard normal distribution:

Density function (bell shaped curve):

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2})(-\infty < z < \infty)$$

Especially of interest is the area under the curve (distribution function), which corresponds for every given value z to the probability of an observation to be smaller than z.

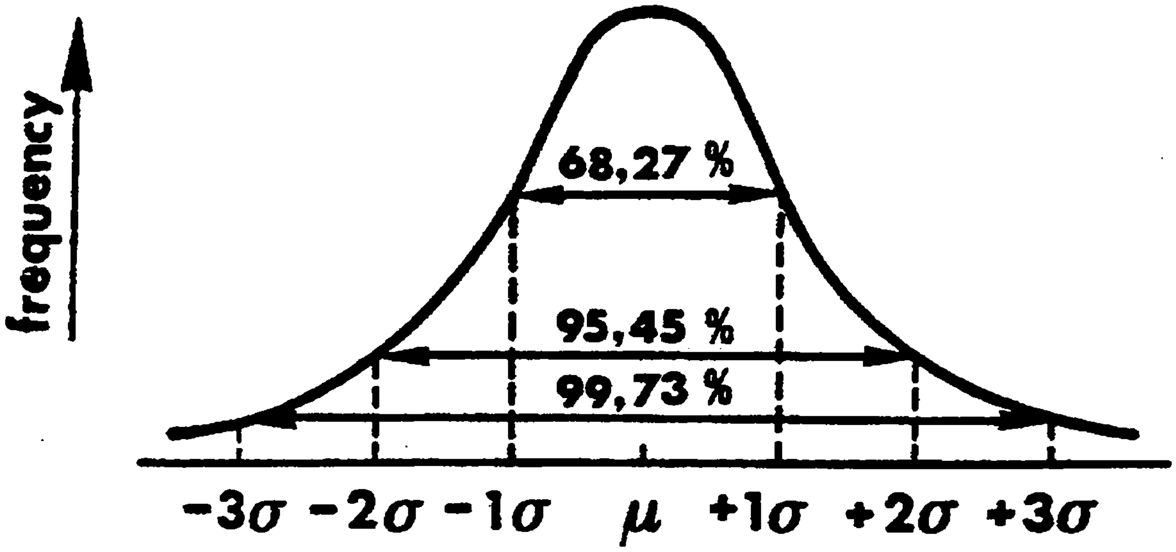
$$\phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-\frac{z^2}{2}) dz, \phi(-z) = 1 - \phi(z)$$



Numerical values for z and ϕ(z) are given in Table A 1, Appendix III. The probability to observe a measured value between two given limits T_1 and T_2 can be calculated as follows:

1) transform the limits T_1 and T_2 in a standardised form
2) calculate the probability with help of Table A 1 by
   $F(z_1, z_2) = \phi(z_2) - \phi(z_1)$

Interpretation of the standard deviation in a normal distribution:

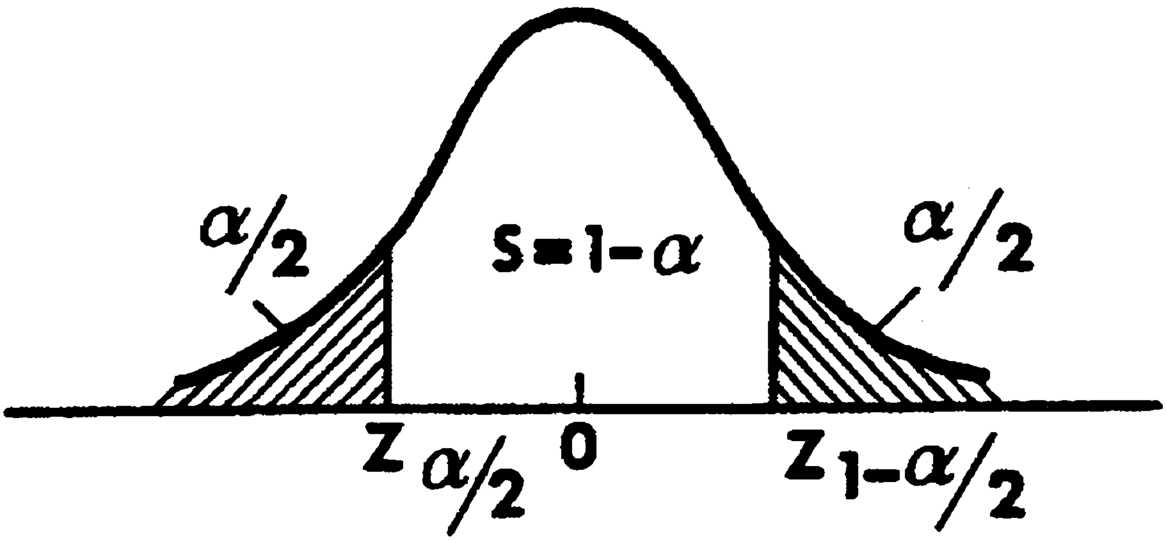


We expect about 95 % of the observation between μ-2σ and μ+2σ . More general we have:

The statement that the values (of the population) lie between $\mu+z_{1-\alpha/2..}$ and $\mu-z_{1-\alpha/2..}$ is right with the probability S = 1 - α and wrong with probability α. One sided or two sided regions may be considered. The corresponding z-values are taken from a table of the standard normal distribution.

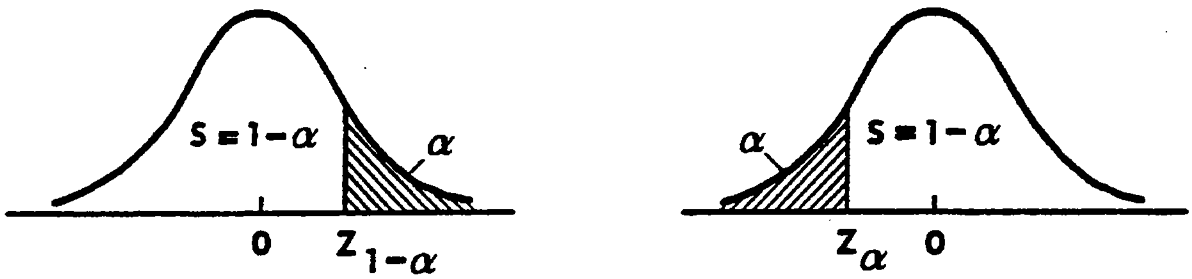
Usual percentiles for the statistical confidence S:

a) Two sided, $S = 1-\alpha; z_{\alpha 2}, z_{1-(\alpha/2)}$



S(%)	90	95	99	99.9
$-z\ \alpha/2 = z_{1}.-\ \alpha/2$	1.64	1.96	2.58	3.29

b) One sided, $S = 1-\alpha; z_{\alpha}, z_{1-\alpha)}$



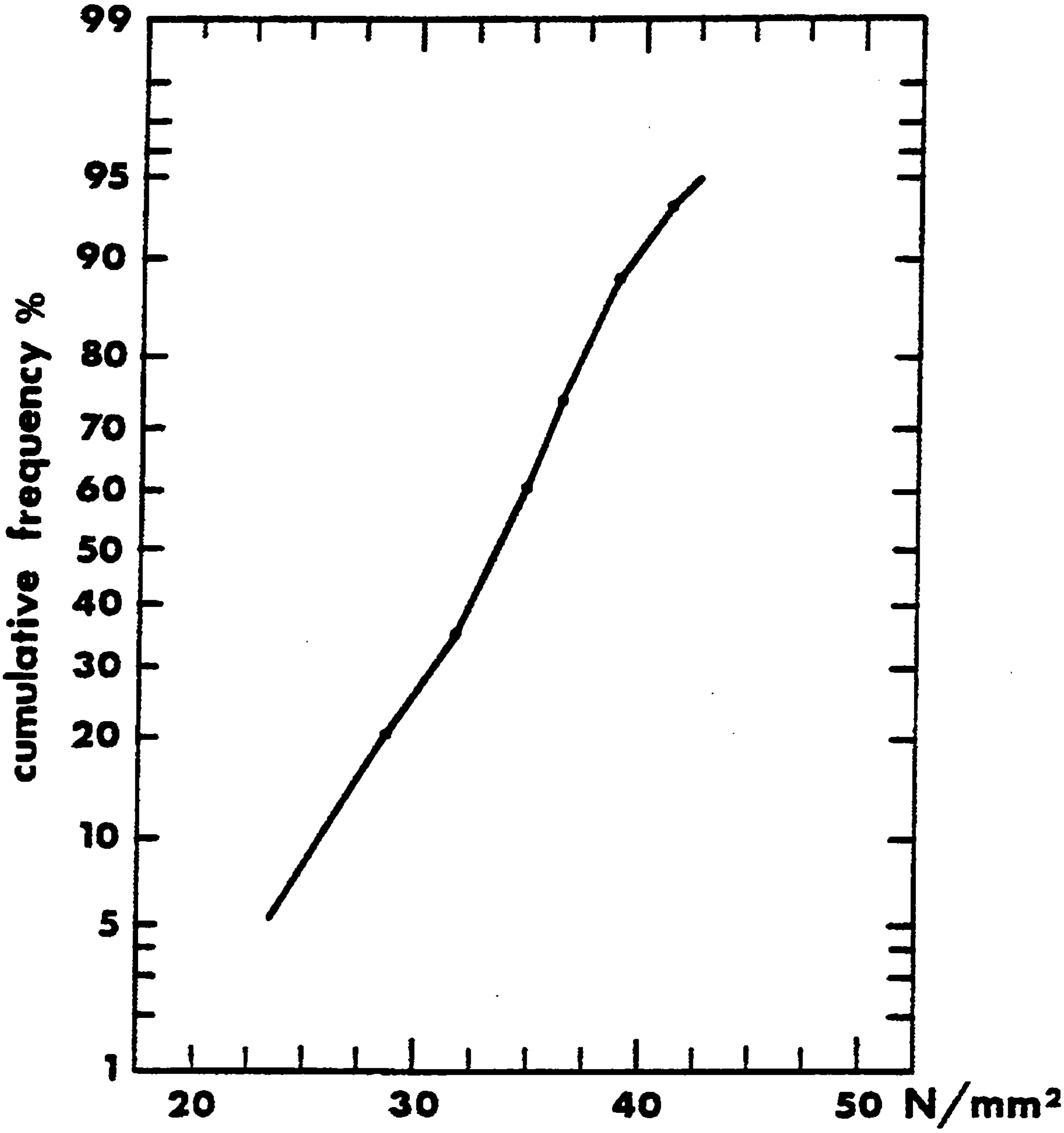
S(%)	90	95	99	99.9
$-z_{\alpha} = z_{1-\alpha}$	1.28	1.64	2.33	3.09

How to check normality?

Before using the characteristics of the ND the validity of the model has to be checked.

Method: Draw the cumulative frequency curve in the normal probability paper (Appendix II). Normality can be assumed if the resulting curve is approximately a straight line between 5% and 95%.

Example: Normal probability plot of compressive strength data



Conclusion: The observations of compressive strength follow approximately a normal distribution.

Short cut rule for rejecting normality: If no negative values are allowed in the observations and the coefficient of variation is greater than 30 %, the distribution is not normal.

4. CONFIDENCE LIMITS

4.1 Confidence limits for the mean

The arithmetic mean is an estimate for the true, unknown mean of the population of all possible observations. We may be interested in the precision of this estimation. Therefore we calculate a confidence interval that contains the true value with high probability (confidence level).

a) If the distribution is normal:

Two sided confidence interval at confidence level 1-α

$$\bar{x} \pm t_{(1-\alpha/2;f)} \cdot \frac{s}{\sqrt{n}}$$

with

f = n-1 degrees of freedom

$t_{1-\alpha/2;f}$ given in Table A 4, Appendix III (Student t-Distribution)

Interpretation:

With probability 1-α the true mean μ lies between the two confidence limits.

For n>50 $t_{1-\alpha/2;f}$ may be replaced by $z_{1-\alpha/2}$ given on page 15 (corresponding to normal deviates of Table A 1, Appendix III). In our example of compressive strength we calculate the 95% confidence interval by replacing the t- by the z- value:

$$\bar{x} \pm t_{0.975;89} \cdot \frac{s}{\sqrt{n}} \cong \bar{x} \pm z_{0.975} \frac{s}{\sqrt{n}}$$

$$\rightarrow 33.2 \pm 1.96 \frac{5.6}{90} = 33.2 \pm 1.16$$

b) If the distribution is not normal:

Approximate confidence intervals can be obtained by

$$\bar{x} \pm z_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

This approximation is derived from the central limit theorem and can be used if

$$n > \frac{10s^2}{\bar{x}^2}$$

c) Range method:

In the case of a normal distribution, confidence limits may be calculated by use of the range instead of the standard deviation (often used in quality control for small sample sizes n).

$$\bar{x} \pm \lambda_{1-\alpha/2} R$$

Values for are given in Table A 6, Appendix III.

4.2 Confidence limits for the median

Confidence limits for the median can be determined directly from the ordered array of observations.

$$x_{(1)} \leq x_{(1)} \leq .... \leq x_{(n-1)} \leq x_{(n)}$$

(1-α)-confidence interval:

$$\underline{x_{(k)} < \text{Median} < x_{(n-k+1)}}$$

with $k = \frac{n-1}{2} - \frac{z_{1-\alpha/2}}{2}\sqrt{n-1}$ rounded to the next lower integer value

$z_{1-\alpha/2}$ is given in Table A 1, Appendix III.

4.3 Confidence limits for the standard deviation

If the population follows a normal distribution, two sided confidence limits are given by

$$s\frac{f}{\chi^2_{1-\alpha/2;f}} \leq \sigma \leq s\frac{f}{\chi^2_{\alpha/2;f}}$$

with f = n 1 degrees of freedom

$\chi^2_{1-\alpha/2;f}, \chi^2_{\alpha/2;f}$ critical values of the chi-square distribution given in Table A 5, Appendix III

Confidence intervals are reduced with increasing sample size n, i.e. the more observations available, the better is the estimation.

The degree of improvement for the arithmetic mean can be derived from the fundamental central limit theorem. If several samples of size n are drawn from the same population, the arithmetic means of these samples are approximately normal with mean μ and standard deviation $s/\sqrt{n}$, the so called standard error of the mean (SEM).

For increasing n to infinity, the standard error of the mean tends to zero, i.e. the estimation tends to be absolutely precise if no systematic errors are present.

4.4 Other methods for the construction of confidence limits

Sometimes it happens that we have to deal with very complicated functions of random variables, for which we can't derive or know the underlying distribution function, but we would like to have confidence limits for the values of this complicated function. A newer method, the Bootshap method (B. Efron, 1983), allows to obtain such results, but is rather intensive in calculation.

5. STANDARD TESTS

5.1 General Test Idea

Tests are used to make decisions in the case of incomplete information. We want to know if some observed difference is significant or only an effect of random errors.

Significant:

We decide that a difference really exists. The probability of this decision to be false is known and can be chosen by the decision maker. A usual choice of this error probability is 5%.

Not significant:

The observed difference may be realized by chance alone. The data give no argument to suppose the existence of a real difference. If such a difference really exists the sample size n is too small to detect it.

Example 3:

At a cement plant A, 618 titration samples gave the mean value $\bar{x}_1 = 82.67$ and the calculated standard deviation $s_1 = 4.13$.

At a later stage a further series of 525 samples were taken and after processing, yielded the information of calculated mean value $x_2 = 85.58$ and calculated standard deviation $s_2 = 3.79$

Your decision problem is:

State whether a significant change has occurred under plant conditions!

Example 4:

The production rate of a cement mill in tons/hr was measured as:

28.3	27.2	29.3	26.7	29.9	24.6	25.0	30.0	26.3	27.8

After modification, the production rate of the same mill in tons/hr was measured as:

28.0	30.0	30.5	26.0	31.0	30.3	24.6	25.4	26.7	29.3

Your decision problem is:

Has the modification made a significant improvement?

There are many problems in which we are interested in whether the mean (or another parameter value) exceeds a given number, is less than a given number, falls into a certain interval, etc.

Instead of estimating exactly the value of the mean (or another parameter), we thus want to decide whether a statement concerning the mean (or other parameter value) is true or false, i.e. we want to test a hypothesis H_0 about the mean (or another value).

In example 3 the hypothesis H_0 could be
"No significant change has occurred in the plant conditions"

In example 4 the hypothesis H_0 could be
"No significant improvement has been made by the modification"

We will solve the two given decision problems later in section 5.2 (test procedures).

Now we will consider another specific example introducing at the same time the important parts of all similar decision problems.

Example 5:

In the manufacture of safety razor blades the width is obviously important. Some variation in dimension must be expected due to a large number of s mall causes affecting the production process. But even so the average width should meet a certain specification. Suppose that the production process for a particular brand of razor blades has been geared to produce a mean width of 0.700 inches. Production has been underway for some time since the cutting and honing machines were set for the last time, and the production manager wishes to know whether the mean width is still 0.700 inches, as intended.

We call set of all blades coming from the production line in a certain time interval (t, t+h) the statistical population to be studied. For example t = January 5th, 0^{00} and t+h = January 6th, 0^{00}, if the statistical population we are interested in is the set of all blades produced on January 5th.

If the production process was initially set up on that day to give a mean width of 0.700 inches, we can say that the hypothesis H_0 "the produced mean width of the regarded population is 0.700" should be tested. In symbols this is: $\mu_0 = 0.700$ = hypothesized mean.

Accepting the Hypothesis H_0:

Suppose we draw a simple random sample of 100 blades from the production line. We measure each of these carefully and find the mean width of the sample to be 0.7005 inches. The standard deviation in the sample turns out to be 0.010 inches. That is,

$n = 100$

$\overline{x} = 0.7005$ inches

$s = 0.010$ inches

For the hypothesis $\mu_0 = 0.700$ to be true, the sample mean $\overline{x} = 0.7005$ inches would have to be drawn from the sampling distribution of all possible sample means whose overall mean is 0.700 inches.

Now the important question arises: If the true mean of the population really were 0.700 inches, how likely is it that we would draw a random sample of 100 blades and find their mean width to be as far away as 0.7005 inches or farther? In other words, what is the probability that a value could differ by 0.0005 inches or more from the population mean by chance alone?

If this is a high probability, we can accept the hypothesis that true mean is 0.700 inches, because it is very easy to get it (high probability).

If the probability is low, however, the truth of the hypothesis becomes questionable because the sample we got is in reality very seldom.

To get at this question, compute the standard error of the mean from the sample:

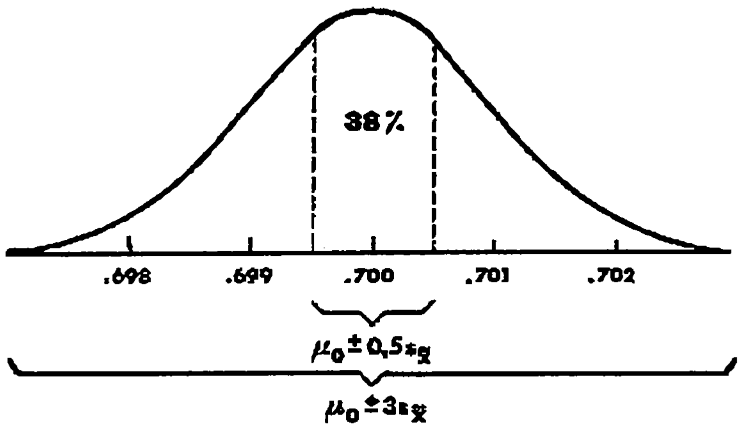
$$s_{\overline{x}} = \frac{s}{\sqrt{n}} = \frac{0.010}{\sqrt{100}} = 0.001 \textit{inches}$$

Since the difference between the hypothetical mean and the observed sample mean is 0.0005 inches and the standard error of the mean is 0.001 inches, the difference equals 0.5 standard errors. By consulting Table A-1 ($z_p = 0.5$)* we find that the area within this interval around the mean of a normal curve is 38%, so that 100 - 38 = 62 % of the total area falls outside this interval (cf. dashed lines below).If 0.700 inches were the true mean, therefore, we should nevertheless expect to find that about 62% of all such possible means would, by chance alone, fall as far away as 0.5 s_x or farther.

Therefore, the probability is 62% that our particular sample mean could fall this far away. This is a substantial reason to accept the hypothesis and attribute to mere chance the appearance of a 0.7005 inches mean in a single random sample of 100 blades.

Naturally we reject in the same time the contrary of H_0 namely that:

$$\mu_0 : \mu \neq 0.700$$



In case H_0 can not be rejected, avoid saying "it is proved that H_0 is correct". or "there are no differences in the means", say "there is no evidence that H0 is not true" or "there is no evidence H_0 should be rejected".

Rejecting the Hypothesis H_0:

Later, after production has gone on for some time, the query again arises:

Is it reasonable to believe that the true mean width of blades produced remains 0.700 inches? Since the process was adjusted to yield that figure the hypothesis still seems reasonable. We could then test it by taking another random sample of 100 blades.

This time the standard deviation is still 0.010 inches, so the standard error of the mean is still 0.001 inches, but the mean is now 0.703 inches:

In order to test the hypothesis that the true mean of the population is 0.700 inches, we again go through the same line of reasoning. If the true population mean really were 0.700 inches, how likely is it that we should draw a random sample of 100 blades and find their sample mean to be as far away as 0.703 inches?

Since the difference between the hypothetical mean of 0.700 inches and the actual sample mean of 0.703 inches is 0.003 inches, and the standard error of the mean is 0.001 inches, the difference is equal to three standard errors of the mean (i.e. 0.003/0.001 = 3).

* $z_p = 0.5$ p = 0.69 = 69%. Therefore 31% fall outside to the right. The same portional fall outside to the left of $-z_p$

Now if 0.700 inches really were the population mean, we know from Table A-1, Appendix III that 99.7% of all possible sample means, for random samples of 100, would fall within three standard errors around 0.700 inches. Hence, the probability is only 0.3% that we would get a sample mean falling as far away as ours does.

We have two choices:

1) We may continue to accept the hypothesis (i.e. leave the production process alone), and attribute the deviation of the sample mean to chance.
2) We may reject the hypothesis as being inconsistent with the evidence found in the sample (hence, correct the production process).

Either of two things is true and we have to make a decision between them:

1) the hypothesis is correct, and an exceedingly unlikely event has occurred by chance alone (one which would be expected to happen only 3 out of 1000 times); or
2) the hypothesis is wrong

Type I and Type II Errors

Understandably, the question can be raised: What critical value should we select for the probability of getting the observed difference ($x - \mu_0$) by chance, above which we should accept the hypothesis H0 and below which we should reject it? This value is called the

critical probability or level of significance

The answer to this question is not simple, but to explore it will throw further light on the nature and logic of statistical decision making. Let's study the following example:

in H0 is true H0 is false

In reality We decide	H_0 is true	H_0 is false
Accept H_0	3 right decision accept a true hypothesis	2 error II accept a false hypothesis
Reject H_0	1 error I reject a true hypothesis	4 error II accept a false hypothesis

Example:

H_0: the parliament building burns

Decision maker is the commander of the fire brigade.

Another expression for error I is: error of first kind

Another expression for error II is: error of second kind

If we ask here, what is worse

- error I or error II,

then error I naturally costs a lot of money because the fire will destroy the whole parliament building. Error II only moves the fire-brigade.

In a long run of cases which the hypothesis is in fact true (although we do not know it is true, for otherwise there would be no need to test it), we will necessarily either be wrong as in 1 or right as in 3.

That is to say, if we make an error it will be of Type I.

Suppose we should adopt 5% as the critical probability. accepting the hypothesis when the probability of getting the observed difference by chance exceeds 5% and rejecting the hypothesis when this probability proves to be less than 5%. This amounts to the decision to accept the hypothesis when the discrepancy of the sample mean is less than 1.96 standard deviations, and to reject the hypothesis when the discrepancy is more than 1.96 standard deviations.

If we take 1% instead of 5%, as above, we will get as limit 2.58 standard deviations.

In fact, the percentage of cases in which we would expect to make an error of the first kind is precisely equal to the critical probability adopted.

(The probability of error I will be abbreviated quite often by α).

Just significant probability level:

In many studies the critical probability is used to describe the statistical significance of a sample result. For example, an economist collects some data on, say, interest rates and the demand for money. He hypothesizes some relationship and wishes to see if the data support his thesis. He tests the hypothesis to rule out the alternative that the observed relationship occurred by pure chance. He reports his sample results as "significant at the 1 percent level". Such a statement is a report to the reader that has the following meaning:

1) if we were to set up a statistical hypothesis
2) if we were to test this hypothesis using a critical probability of 1%
3) then we would reject the hypothesis and rule out a chance relationship

Significance levels of 10%, 5%, 1%, 0.1% are often used in reporting sample data. The smallest of these probability values is chosen at which the hypothesis can be rejected.

So we see now what is basic for every statistical test:

1) we need a clear hypothesis H_0
2) we have to know by which statistic we want to test H_0 (in our example it was $\overline{x}$)
3) we need a criterion C for decision making in the following form:
   reject H_0 if C applies
   accept H_0 if C does not apply

The criterion C is usually given in form of a critical limit, called significance limit, which should not be exceeded by the calculated test statistic.

To perform such a statistical test, it is necessary to select the risk to commit a type I error, i.e. the significance level α (see section 5.2, p. 28).

Often a relevant difference has to be detected with a certain probability, i.e. with a predetermined type II error β. This is only possible with a certain sample size n, as is further outlined in section 5.3.

If we reject a hypothesis when it should be accepted, we say that a Type I error has been made. If, on the other hand, we accept a hypothesis when it should be rejected, we say that a Type II error has been made. In either case a wrong decision or error in judgement has occurred.

In order for any tests of hypotheses or rules of decision to be good, they must be designed so as to minimize errors of decision. This is not a simple matter since, for a given sample size, an attempt to decrease one type of error increases the other type. In practice one type of error may be more serious than the other, and so a compromise should be reached in favour of a limitation of the more serious error. The only way to reduce both types of error is to increase the sample size, which may or may not be possible.

Note that from the philosophy of testing there is always the possibility to reject H_0 although H_0 effectively is true (the probability for this is α, the type I error probability). So if you are testing say 100 "correct" datasets (i.e. for which H_0 holds) to a significance level $\alpha = 5\%$ you will expect about five results that reject H_0, although it holds. This has to be considered if a lot of (statistical) tests are made on the same data material (see e.g. Multiple Testing ...."Simultaneous Statistical Inference", Miller, 1981).

One-sided and two-sided tests

In example 5 we are interested in values of significantly higher or smaller than μ_0. Any such test which takes account of departures from the null hypothesis H_0 in both directions is called a two-sided test (or two-tailed test $H_0 : \mu = \mu_0, A : \mu \neq \mu_0$)

However, other situations exist in which departures form H_0 in only one direction is of interest. In example 4 we are interested only in an improvement of production rate due to the modification of the cement mill and so a one-sided test is appropriate. It is important that the decision maker should decide if a one-tailed or two-tailed test is required before the observations are taken.

5.2 Test Procedures

Performance of a statistical test requires in advance the recognition and formulation of the present decision problem (test situation). Use the following checklist for applications of statistical tests:

CHECKLIST

1) Is the decision problem concerned with the mean. median. standard deviation. correlation coefficient or other?
   Formulate a clear hypothesis H_0.
2) Is it a one-sample or a two-sample problem?
   One-sample problem:
   Mean, median or standard deviation of a given sample of size n is compared with a hypothetical value (example 5).
   Two-sample problem:
   The decision problem is concerned with differences between two given samples.
3) For the two-sample problem: Are the observations independent or paired? Observations are paired if every value in the first sample can be attached definitely to a value in the other sample. Paired comparisons are in general considerably more efficient than a comparison of independent samples.
4) Is the sample drawn from a normal distribution?
5) Decide between one-sided or two-sided test.
   One-sided tests are used if differences only in one direction may occur or are of interest. Two-sided tests are used if no preliminary information is available, in what direction the sample may differ.
6) Choose the test procedure in the table "TEST CONCERNED WITH" (sample sizes n<50 are considered to be small).
7) Choose the significance level a. Usual choices are $\alpha = 5\%$ or $\alpha = 1\%$ (error of type I).
8) Calculate the test statistic T of the chosen test (cf. the formulae of the following pages).
9) Look for the significance limit T_p for the test statistic in the corresponding table with $p = 1-\alpha/2$ (two-sided test) or $p = 1-\alpha$ (one-sided test).
10) Decision: If the calculated test statistic exceeds the significance limit, reject the hypothesis H_0 (the observed difference is significant at the level α). Otherwise there is no reason to reject the hypothesis H_0 and the difference is considered to be not significant.

TEST CONCERNED WITH

test situation	**median**	**mean general**	**normal distribution**	**mean standard deviation**
One-sample problem				
small n (n<50)	sign test	signed-rank-test	t-test	x^2-test
large n (n≥50)	sign test	z-test	z-test	x^2-test
Two-sample problem independent				
small n	median test	Wilcoxon-test	t-test	F-test
large n	median test	z-test	z-test	F-test
paired				
small n	sign test	signed-rank-test	t-test	-
large n	sign test	z-test	z-test	-
More than two samples	x^2-test	Krukskal- Wallis-test	Analysis of variance (ANOVA) Friedmann-test	Bartlett-test
(These tests are not treated i n this paper)				

Selected test statistics

5.2.1 z-Test

Used to test means with large sample size n.

a) One-sample-problem:
Given is a sample of size n with arithmetic mean $\overline{x}$ and standard deviation s, we test the hypothesis that the mean (estimated by $\overline{x}$) is equal to a given or target value μ_0
Hypothesis H_0 $\mu = \mu_0$

Test statistic $$z = \frac{(\overline{x} - \mu_0)\sqrt{n}}{\sigma}$$

In general σ is not known. It can be replaced by s for large sample sizes.

b) Two independent samples:
Given are two samples of size n_1 and n_2 with arithmetic means $\overline{x}_1, \overline{x}_2$ and standard deviations s_1, s_2.
Hypothesis H_0 $\mu_1 = \mu_2$

Test statistic $$z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

c) Paired comparison
Given is a sample of n paired observations x_i, y_j. Calculate the arithmetic mean $\overline{d}$ of all the differences $d_i = y_i - x_i$ and the standard deviation s_d.
Hypothesis H_0 $\mu_x = \mu_y$ respectively $\mu_d = 0$

Test statistic $$z = \frac{\overline{d}\sqrt{n}}{s_d}$$

Significance limits

Often used significance limits for $z_{1-\alpha}$ (one-sided test) and $z_{1-\alpha/2}$ (two-sided test) are

$z_{0.95} = 1.645$ $\quad z_{0.975} = 1.960$

$z_{0.99} = 2.326$ $\quad z_{0.995} = 2.576$

For other significance levels α look for $z_{1-\alpha}$ in Table A-1.

Decision

The difference is significant if

$$|z| \geq z_{1-\alpha/2} \quad \text{two sided test}$$

$$\left.\begin{array}{l} |z| \geq z_{1-\alpha} \\ |z| \leq -z_{1-\alpha} \end{array}\right\} \quad \text{one-sided test}$$

Example

a) In the example 1 we test the two-sided hypothesis H_0:
$\mu = 35.0$ N/mm^2 $= \mu_0$ (one-sample problem).
$\bar{x} = 33.2$
$s = 5.6$
$n = 90$
$\alpha = 0.05$

$$z = \frac{(33.2 - 35.0)\sqrt{90}}{5.6} = -3.05$$

$|z| = 3.05$ ' $1.96 = z_{0.975}$
Decision: The sample mean differs significantly from the standard strength 35.0 N/mm^2.

b) in example 3 (chapter 5.1) we test the hypothesis whether a change in plant conditions has occurred or not (two independent samples):
H_0: $\mu_1 = \mu_2$
$n_1 = 618$, $\quad n_2 = 525$
$x_1 = 82.67$, $\quad x_2 = 85.58$
$s_1 = 4.13$, $\quad s_2 = 3.79$
$\alpha = 0.05$ (two-sided test)
85.58-82.67

$$z = \frac{85.58 - 82.67}{\sqrt{\frac{(4.13)^2}{618} + \frac{(3.79)^2}{525}}}$$

$$z = 12.41 > 1.96 = z_{0.975}$$

Decision: Highly significant difference between the two samples of titration values.

5.2.2 Sign Test

Test for the median in one-sample problems and paired comparisons.

a) One-sample problem: Given is a sample of size n.
We test the hypothesis that the median of the population is equal to a given value m_0
H_0: median = m_0
Test statistic:
Count the number of observations smaller than m_0 and those larger than m_0. The test statistic k is then the smaller of the two numbers.

b) Paired comparison: Given are n paired observations x_i, y_j.
We test the hypothesis that x_i and y_j have the same median.
Test statistic:
For each pair of observations x_i, y_j record the sign of the difference y_i - x_i .
The test statistic k is then the number of occurrences of the less frequent sign.

Significance limits

Calculate $k_{\alpha/2} = \frac{n-1}{2} - \frac{z_{1-\alpha/2}}{2}\sqrt{n-1}$ (two-sided test).

For the one-sided test replace $\alpha/2$ by α. Z_p is the significance limit of the z-test

$$z_{0.95} = 1.645 \qquad z_{0.975} = 1.960$$

Decision

If k is less than $k_{\alpha/2}$ (resp. k_α) conclude that the medians are different, otherwise, there is no reason to believe that the medians differ. 44

Example

Test the hypothesis that 50% of the population has an income of more than 3000 (example 2).

H_0: m_0 = 3000
k = 217
n = 479
α = 0.05 two-sided test

$$k_{0.025} = \frac{478}{2} - \frac{1.96}{2}\sqrt{478} = 217.6$$

The test is just significant at the 5% level. Because the sample median is 2700, we conclude that less than 50% of the population has an income of 3000.

5.2.3 Signed-rank test

Test for the mean in symmetrical distributions with small sample sizes n (one-sample or paired comparison).

a) One-sample problem: Given is a sample of size n from a symmetrical distribution.
Hypothesis $H_0 : \mu = \mu_0$

Disregarding signs, rank the d_i according to their numerical value, i.e., assign rank 1 to the smallest observation, assign the rank of 2 to the d_i which is next smallest, etc. In case of ties, assign the average of the ranks which would have been assigned if the d_i's had differed only slightly. (If more than 20% of the observations are involved in ties, this procedure should not be used).

To the assigned ranks 1, 2, 3, etc., prefix a + or a - sign, according to whether the corresponding d_i is positive or negative. The test statistic T is then the sum of these signed ranks.

a) Paired comparisons: Given are n paired observations x_i, y_i.
The hypothesis is tested that both have the same mean.
Test statistic: Compute $d_i = y_i - x_i$ for each pair of observation and continue as in a).

Significance limits

Look for $T_{0.95}$ or $T_{0.975}$ in Table A-7. If the number m of differences d_i exceeds 20 perform a z-test with

$$z_T = \frac{T}{\sqrt{\frac{m(m+1)(2m+1)}{6}}}$$

Decision

Conclude that the means differ if

$\|T\| \geq T_{1-\alpha/2}$	$(or \|z_T\| \geq z_{1-\alpha/2})$	two-sided test
$T \geq T_{1-\alpha}$	$(or\, z_T \geq z_{1-\alpha})$	one-sided test
$T < -T_{1-\alpha}$	$(or\, z_T < -z_{1-\alpha})$	one-sided test

5.2.4 Wilcoxon-Test

Test for a comparison of means in two independent samples:

Given are two independent samples of size n_1 and n_2.

We test the hypothesis that both samples have the same mean

$$H_0: \mu_1 = \mu_2$$

Test statistic:
Combine the observations from the two samples, and rank them in order of increasing size from smallest to largest. Assign the rank of 1 to the lowest, a rank of 2 to the next lowest, etc. (Use algebraic size, i.e., the lowest rank is assigned to the largest negative number, if there are negative numbers). In case of ties, assign to each the average of the ranks which would have been assigned if the tied observations had differed only slightly. (If more than 20% of the observations are involved in ties, this procedure should not be used).

Let n_1 = smaller sample
n_2 = larger sample
$n = n_1 + n_2$

Compute R, the sum of the ranks for the smaller sample. (If the two samples are equal in size, use the sum of the ranks for either sample).

Compute $W = 2R - n_1(n + 1)$

Significance limits

Look for $W_{0.95}$ or $W_{0.975}$ in Table A-8. For sample sizes not mentioned in the table perform a z-test with

$$z_W = \frac{W}{\sqrt{\frac{n_1 n_2 (n+1)}{3}}}$$

Decision

Conclude that the means differ if

$|W| \geq W_{1-\alpha/2} \quad (or |z_w| \geq z_{1-\alpha/2})$ two-sided

$W \geq W_{1-\alpha} \quad (or z_w \geq z_{1-\alpha})$ one-sided

$W < -W_{1-\alpha} \quad (or z_w < -z_{1-\alpha})$ one-sided

Example

In example 4 (chapter 5.1) we are interested in an improvement of production rate after a modification of the cement mill.

$\alpha = 0.05$ (one-sided test)

Combined sample (values in italics correspond to the sample after modification):

tons/hr	<u>24.6</u>	*24.6*	*25.0*	<u>25.4</u>	<u>26.0</u>	*26.3*	26.7	<u>26.7</u>	*27.2*	*27.8*
rank	1.5	*1.5*	*3*	4	5	*6*	7.5	7.5	*9*	*10*
tons/hr	<u>28.0</u>	*28.3*	<u>29.3</u>	*29.3*	*29.9*	<u>30.0</u>	<u>30.0</u>	<u>30.3</u>	<u>30.5</u>	<u>31.0</u>
rank	11	*12*	13	*14*	*15*	16.5	16.5	18	19	20

The sum R of ranks in the second sample is
R = 1.5 + 4 + 5 + 7.5 + 11 + 13 + 16.5 + 18 + 19 + 20 = 115.5
W = 2 . 115.5 - 10 . 21 = 21
$W = 21 < 46 = W_{0.95}$

Decision

The improvement is not significant. We have no reason to assume a real improvement due to the modification.

5.2.5 t-Test

> This test is used for comparisons of means in normal distribution with small sample sizes n. If the distribution is not known to be normal, prefer rank tests (Wilcoxon, signed ranks).

The test situation is the same as in 5.2.1 (z-test) refer for remarks on hypothesis and decision. Instead of $z_{1-\alpha}$ look for significance limits $t_{1-\alpha}$ in Table A-4, with n-1 degrees of freedom (df) in the one-sample case or for paired comparison, and with n1 + n2 - 2 degrees of freedom (df) for the two-sample case respectively.

One-sample problem: Given is a sample of size n drawn from a normal distribution.

$$t=\frac{(\bar{x}-\mu_0\sqrt{n}}{s}$$

Two independent samples: Given are two independent samples of size n_1 and n_2 from two normal distributions with means μ_1, μ_2 and equal standard deviation σ.

$$t=\frac{\bar{x}_2-\bar{x}_1}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}}}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$$

Paired comparison: Given are n pairs of observations x_i, y_i, when x and y follow a normal distribution.

$$t=\frac{\bar{d}}{s_d}\sqrt{n}$$

5.2.6 Median-Tests

Median tests for two independent samples are not given here in an explicit form. For small sample sizes n Fisher's test for 2 x 2 contingency tables and for large n x^2-test for contingency tables may be used (Natrella 1963, Noether 1971).

In a comparison of two independent samples we are often not interested in differences only of means or medians, but generally in location differences of the two samples.

In this case use the tests for means, which are more efficient than median tests (Wilcoxon-test or z-test).

5.2.7 X^2-Test

Several X^2-tests are available for different test situations. The test we explain here is used to compare the standard deviation s of a sample with hypothetical value σ_0.

X^2-tests are also used for

- Goodness of fit test
- Independence test
- Loglinear models

(see e.g. Haberman, 1978)

Given is a sample of size n drawn from a normal distribution with mean μ and standard deviation σ. We test the hypothesis that the standard deviation σ, estimated by s from the sample is equal to a hypothetical standard deviation σ_0.

$$H_0: \sigma = \sigma_0 \quad \text{Test statistic: } x^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

Significance limits

Look for significance limit $X^2_{p,m,}$ in Table A-6 with m = n-1 degrees of freedom (df).

Decision

If	$X^2 \geq X^2_{1-\alpha/2;m}$	or $X^2 \leq X^2_{\alpha/2;m}$	conclude: $\sigma = \sigma_0$	(two-sided)
If	$X^2 \geq X^2_{1-\alpha;m}$		conclude: $\sigma > \sigma_0$	(one-sided)
	$X^2 \leq X^2_{\alpha;m}$		conclude: $\sigma < \sigma_0$	(one-sided)

Otherwise we have no reason to believe that σ differs from σ_0.

5.2.8 F-Test

Comparison of the standard deviation in two independent samples.

Given are two independent samples of size n_1 and n_2 with standard deviations s_1 and s_2. Both samples are drawn from a normal distribution. We test the hypothesis that both populations have the same standard deviation.

$$H_0: \sigma_1 = \sigma_2$$

Test statistic:

Let be $s_1 > s_2$, then compute

$$F = \frac{s_1^2}{s_2^2}$$

Significance limits

Look for $F_{1-\alpha;m1,m2}$ (one-sided) or $F_{1-\alpha/2;m1,m2}$ (two-sided) in Table A-9 with $m_1 = n_1-1$ and $m_2 = n_2-1$ degrees of freedom.

Decision

Conclude $\sigma_1 \neq \sigma_2$ if $F \geq F_{1-\alpha/2;m1,m2}$ (two-sided)

or $\sigma_1 > \sigma_2$ if $F \geq F_{1-\alpha;m1,m2}$ (one-sided)

5.3 Sample Size Determination

As stated at the end of chapter 5.1, the probability β of making a type II error in a given test with significance level α (type I error) depends upon the sample size n.

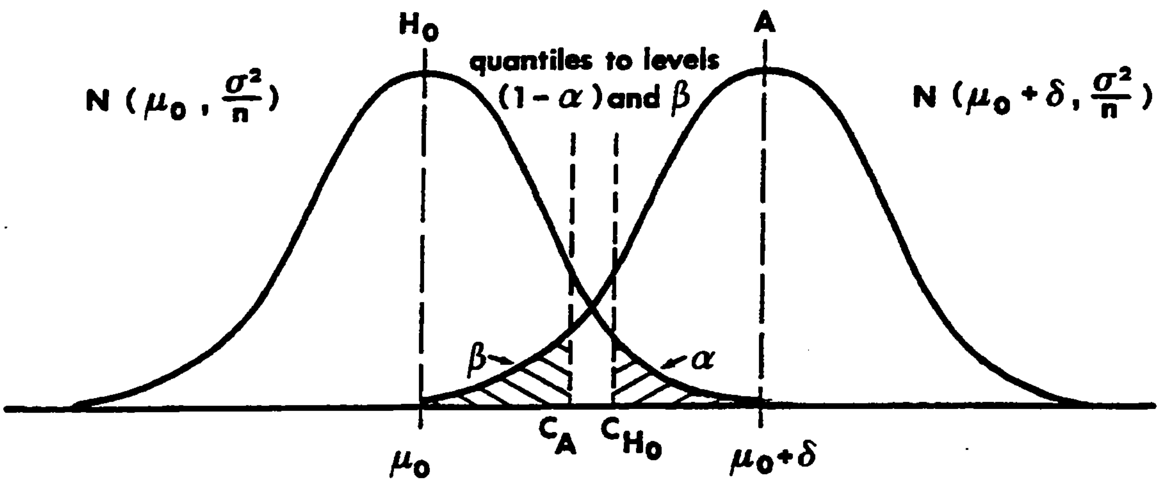
How can we determine the sample size that is necessary to hold the type II error within certain boundaries (probability β)? Or in other words: What sample size n is necessary to detect a relevant difference with great probability (1-β)?

Let us first examine a one-sided one-sample-test for testing a mean:

Suppose we are interested in the mean μ of a random variable X. We want to test the hypothesis H_0: $\mu = \mu_0$ against the alternative hypothesis A: $\mu > \mu_0$ with a z-test. The error-type-one shall be α.

The probability to detect a deviation of Δ from μ_0 shall be at least (1 - β), if this deviation exceeds a preselected relevant difference δ (β is now an upper bound for the type II error).

The following figure shows the distribution of the sample mean $\overline{x}$ under H_0 and under the (special) alternative hypothesis A: $\mu = \mu_0 + \delta$:



Interpretation of the graph:

- C_{H_0} is the criteria (significance limit) of the test:
  The hypothesis H_0 is rejected with probability α even if it is true.
- On the other hand, the probability to accept H_0: $\mu = \mu_0$, even if the true mean is $\mu = \mu_0 + \delta$ (type II error), is greater than β.
- The relevant difference δ may only be detected with probability (1-β), if $C_{H_0} \leq C_A$; in this case the risk to commit a type II error is ≤ β.

The variance $\frac{\sigma^2}{n}$ of the two distributions becomes smaller with increasing n and hence C_A moves to the right and C_{H_0} to the left. Therefore, we can find an n such that $C_{H_0} \leq C_A$:

$$C_{H_0} \leq C_A$$

$$z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}} + \mu_0 \leq z_\beta \frac{\sigma}{\sqrt{n}} + \mu_0 + \delta = - z_{1-\beta} \frac{6}{\sqrt{n}} + \mu_0 + \delta \qquad (1)$$

$$\frac{\sigma}{\delta}(z_{1-\alpha} + z_{1-\beta}) \leq \sqrt{n}$$

or $n \geq \frac{\sigma^2}{\delta^2}(z_{1-\alpha} + z_{1-\beta})^2$ (2)

n corresponds with the required sample size to detect a desired difference δ with probability (1-β) by a z-test with significance level α.

For the two-sided one-sample-test we only have to replace α by α/2, and we get:

$$n \geq \frac{\sigma^2}{\delta^2}(z_{1-\alpha/2} + z_{1-\beta})^2 \qquad (3)$$

Note: β is not replaced by β/2!

In a similar (but somewhat more complicated) way we find lower bounds for the sample sizes n_1 and n_2 satisfying our conditions in the one-sided two-sample-test:

$$n_2 \geq \frac{\sigma^2 \sigma_1 + \sigma_2^2}{\delta^2}(z_{1-\beta} + z_{1-\alpha})^2 \qquad (4)$$

$$n_1 \geq \frac{\sigma_1}{\sigma_2} n_2$$

For equal standard-deviations ($\sigma_1 = \sigma_2 = \sigma$) we obtain from (4):

$$n_1 = n_2 \geq \frac{2\sigma^2}{\delta^2}(z_{1-\beta} + z_{1-\alpha})^2 \qquad (5)$$

In the two-sided two-sample-test we simply have to replace α by α/2 in (4) or (5).

The following table shows short rules for determining sample sizes when α = β = 5% and $\sigma_1 = \sigma_2 = \sigma$ in the two-sample-case. The numbers in parentheses refer to the formulae from which the rules are derived.

	one sample	two samples
one-sided	$n \geq \frac{11\sigma^2}{\delta^2}$ (2)	$n_1 = n_2 \geq \frac{11\sigma^2}{\delta^2}$ (5; for $\sigma_1 \neq \sigma_2$ see (4))
two-sided	$n \geq \frac{13\sigma^2}{\delta^2}$ (3)	$n_1 = n_2 \geq \frac{26\sigma^2}{\delta^2}$ (5)

with δ = relevant difference to be detected with probability (1-β)

α = β = 5%

Note:

- The variance σ^2 is usually not known, but often some knowledge about σ^2 is available from former experiments (standard deviation). Otherwise σ^2 may be estimated in a pilot study.
- In the case of small sample sizes (t-test, Wilcoxon-test), add 5% to the calculated n.

6. DATA PRESENTATION AND INTERPRETATION

(Chambers, Cleveland, Kleiner and Mikey, 1983).

The problem of data presentation and interpretation is common to cement manufacturers and users. In every stage of cement, aggregate, concrete production and application, data concerning materials, equipment, energy consumption, market situation, costs, etc. are generated. All this information must increase the knowledge about what happens in the process and market, thus providing the basis of decision. Moreover, communication between supplier and consumer has to rely on this information. We consider it, therefore, vital not only for optimum manufacture of cement and cement based products, but also with respect to the mutual relation between manufacturer and user that appropriate attention is given to data handling, i.e. presentation and evaluation.

We may differentiate between three levels of information:

- data needed for direct process and quality control (routine decision, off- or on-line)
- data required for day-to-day management decisions based on quality reports
- additional data allowing long-term improvements and developments.

With respect to presentation and interpretation of test results, the main problems for plant management are:

- how to organize the required information flow so that decision can be made at all levels of competence
- to reproduce and evaluate the corresponding data in a specific situation.

The relevant data of al1 domains cannot be made available without a functional information system. But availability alone does not necessarily result in a rational decision based on the data. To achieve this, two further aspects must be considered: first, the data should be presented in an intelligible form, i.e. a high transparency of the results enables the management to recognize certain relationships or critical results in time. Secondly, the decision maker must be aware of accuracy, significance and relevance of the considered data in order to provide a realistic interpretation.

Availability of data

A quick availability of data does not necessarily require a fully integrated data based system with electronic data processing, but is rather a matter of organization. Important data (e.g. for process and quality control) should circulate with little loss of time. Consequently, an appropriate reporting system (organization) must be established. Additional data should not disappear in some drawer where a later retrieval is impossible or at least very inconvenient. To avoid such a disorder, all data are recorded in a similar way, including a note on the circumstances of measurement and provenance of samples and test results. Remarks about circumstances and provenance are used to judge the comparability of different sets of observations. A standardised recording procedure makes data surveying easy and simplifies a later change to electronic data processing.

6.1 Intelligible Presentation

The main aim of data presentation is transparency rather than secrecy, i.e. graphs should be employed instead of tables. A graphical presentation gives a quick survey on relevant information such as changes in time, extreme values, relationship between variables. A short survey on frequently used graphs can be found in section 6.4. It is recommended to produce the graphs directly at the source of the data in the form of tallies and/or control charts. A control chart may already be included in the laboratory journal and reports, directly behind the columns for sample identification and test results. The possibility of obtaining a quick survey by consulting a well presented graph will not only help the management in its decision making, but transparency will also improve interest and motivation of the personnel at any level of competence.

6.2 Interpretation

First of all, we must know how much we can rely on the data.

Test results are never "true" values, but are rather subject to errors of three types (see also section 1):

- random errors caused by sampling, imprecise measurement, environmental effects, etc.
- systematic errors caused by bias in sampling or process measurement, by the use of inadequate experimental design
- gross errors caused by recording wrong or not comparable values.

Usually, we expect 5 to 10% gross errors in a set of observations. Random errors may be denoted as reproducibility and expressed as the corresponding standard deviation. In general, these errors are underestimated. Reproducibility should be known for every important analyzing method. Systematic and gross errors are difficult to characterize. They should be minimized by careful experimentation and data handling. A periodic check for systematic errors should be done (calibration, comparison with a standard, inter-laboratory test).

How to compare several data sets or groups of data?

Before any comparison is made, we have to seriously check the comparability of data. Often, data sets differ in provenance of samples or circumstance of measurement, so that a comparison ma y be impossible. Even the fact that sample 1 was measured by laboratory assistant Miller and sample 2 by Brown will lead to a biased comparison if there is any relevant systematic error between the two persons. If the data sets are comparable and we observe a certain difference, we have to ask the following questions before taking any action:

a) Is the difference significant?

b) Is the difference relevant?

The problem of significance is answered by a statistical test. If it is not significant, we have no reason to take any action because the observed difference may occur by chance alone. If the test indicates a significant difference, it is not necessarily relevant for the problem we are concerned with. Of course, the decision whether or not it is relevant is not a statistical problem. Perhaps a decision maker is alarmed when an observed difference, considered to be relevant in the present problem, does not lead to a significant test result. In this case, the sample size used may be too small or the testing procedures may not be sufficiently sensitive to solve the given problem.

Consequently, experience in product manufacture and application is necessary to assess the relevance of difference between target quality and experimental values. However, to judge whether it is significant requires a training in statistical technique, particularly test procedures. Decisions based simply on either practical experience of many years (relevance) or on statistical technique and procedures (significance) will lead to too frequent and unnecessary actions in both the manufacturing process and product application.

6.3 Data Interpretation related to Problems in Cement Application

If confronted with the problem of finding the reasons for poor quality of cement related products (concrete, asbestos cement, etc.), the cement is often suspected to be the cause. This can be explained by the fact that cement is the binding agent, and in the majority of cases constitutes the most expensive component, although it is known that the effect of other components, proportioning and curing conditions etc. are of great importance. A further reason to inspect the cement first is the difficulty to specify the other effects while cement is well defined by its chemical and mineralogical composition and its fineness.

To find out the causes for inferiority or changes in quality, a collaboration of consumer, producer and statistician is indispensable: the statistician may be omitted in uncomplicated problems if delegates of consumer or producer are well trained in statistics. In a first retrospective analysis of available data, parallel changes of parameters and quality are studied. This is done by drawing scatter diagrams and performing a regression analysis. The disadvantage of retrospective studies is the difficulty to find out causal relationships. The effects of several variables are mixed and cannot be separated due to their causal origin. On the other hand, an observed correlation indicates a possible causal effect. The decision about what cause is really responsible must be made by an experienced specialist and not by a statistical test. Often a decision is not possible because dependencies are too complex. In this case, a special experiment has to be planned and performed with a controlled variation of suspected variables and careful elimination of interfering effects (prospective analysis). In contrast to the retrospective analysis, an adequately designed experiment renders it possible to evaluate and judge causal effects with statistical methods. A short survey on the use and interpretation in regression/correlation analysis and designed experiments is given in chapter 7.

Survey on Graphical Data Representation

Representation of one sample - documentation

Frequency table / Tally

Given is a sample of large size n. Observations are grouped into classes of equal length and marked in the tally.

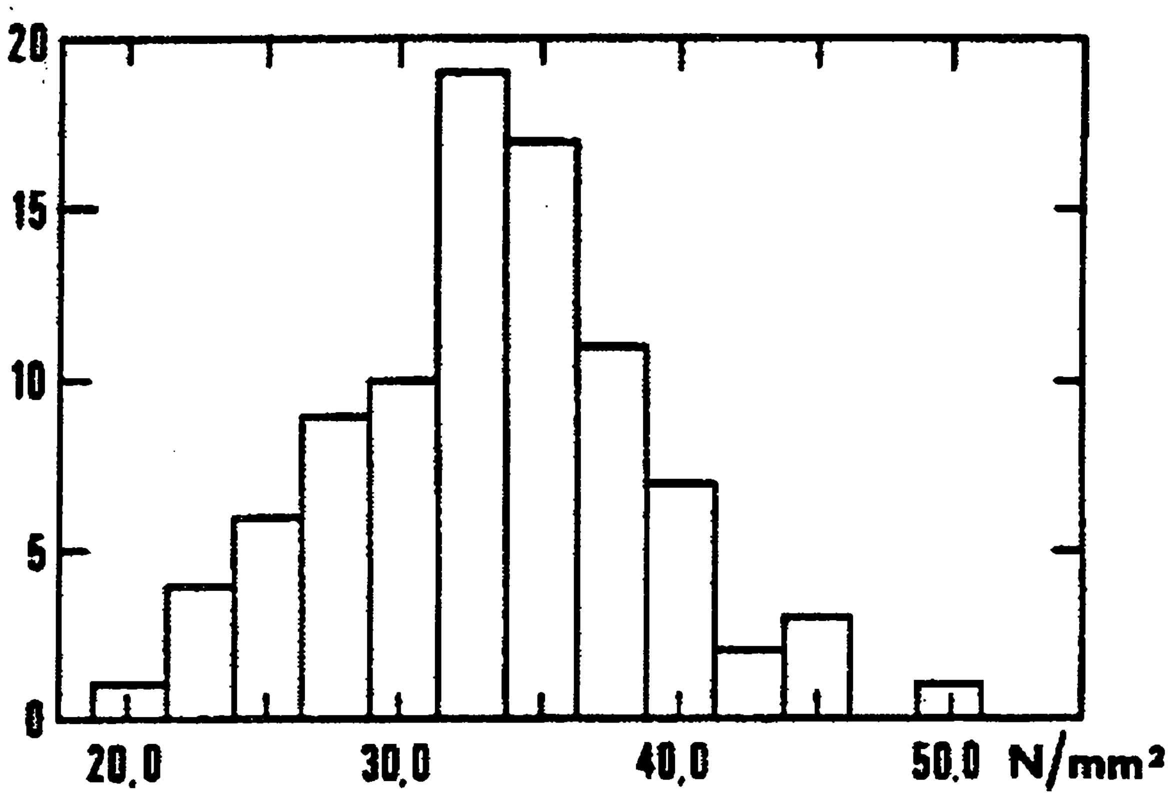
Example: 90 values of concrete strength

mid-point	class	tally	absolute frequency	relative frequency
20.0	18.75 - 21.25	|	1	0.011
22.5	21.25 - 23.75	||||	4	0.044
25.0	23.75 - 26.25	卌 |	6	0.067
27.5	26.25 - 28.75	卌 ||||	9	0.100
30.0	28.75 - 31.25	卌 卌	10	0.111
32.5	31.25 - 33.75	卌 卌 卌 ||||	19	0.212
35.0	33.75 - 36.25	卌 卌 卌 ||	17	0.189
37.5	36.25 - 38.75	卌 卌 |	11	0.122
40.0	38.75 - 41.25	卌 ||	7	0.078
42.5	41.25 - 43.75	||	2	0.022
45.0	43.75 - 46.25	|||	3	0.033
47.5	46.25 - 48.75		0	0.000
50.0	48.75 - 51.25	|	1	0.011
Total			90	1.000

Histogram

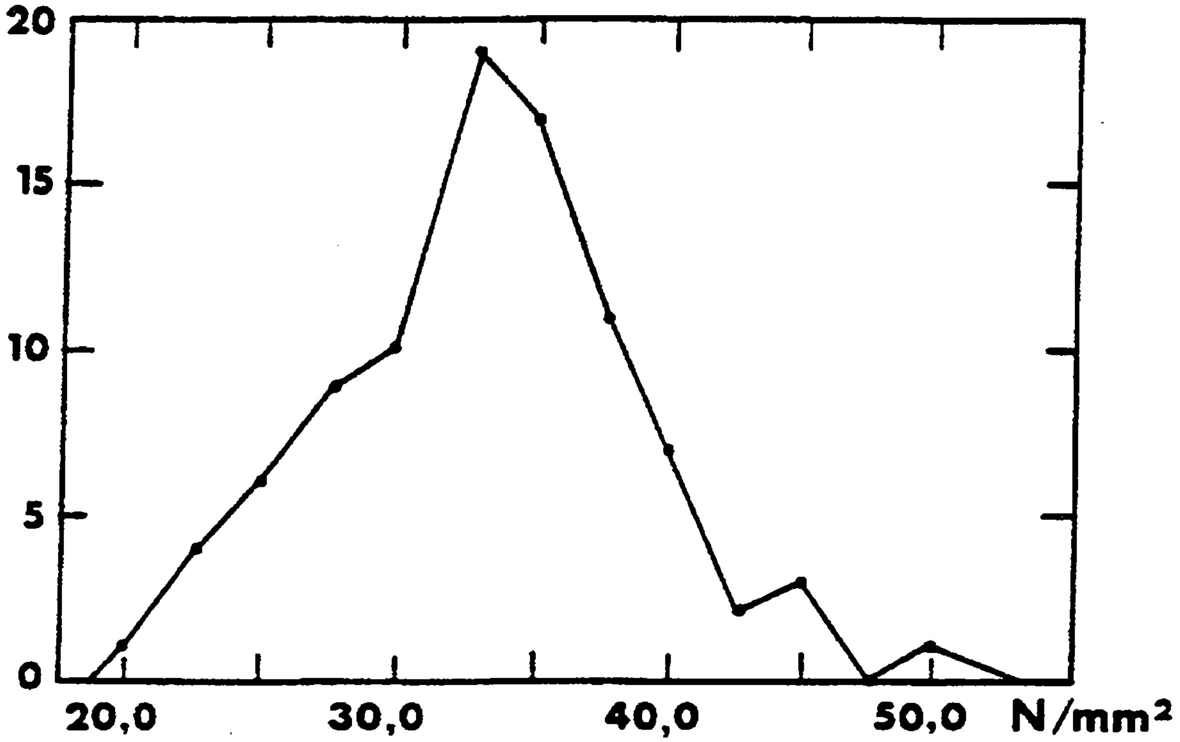
Graphical representation of the tally. Visualization of minimum, maximum, center and shape of the distribution.

Area of rectangles corresponds to absolute or relative frequencies in the classes.



Frequency curve

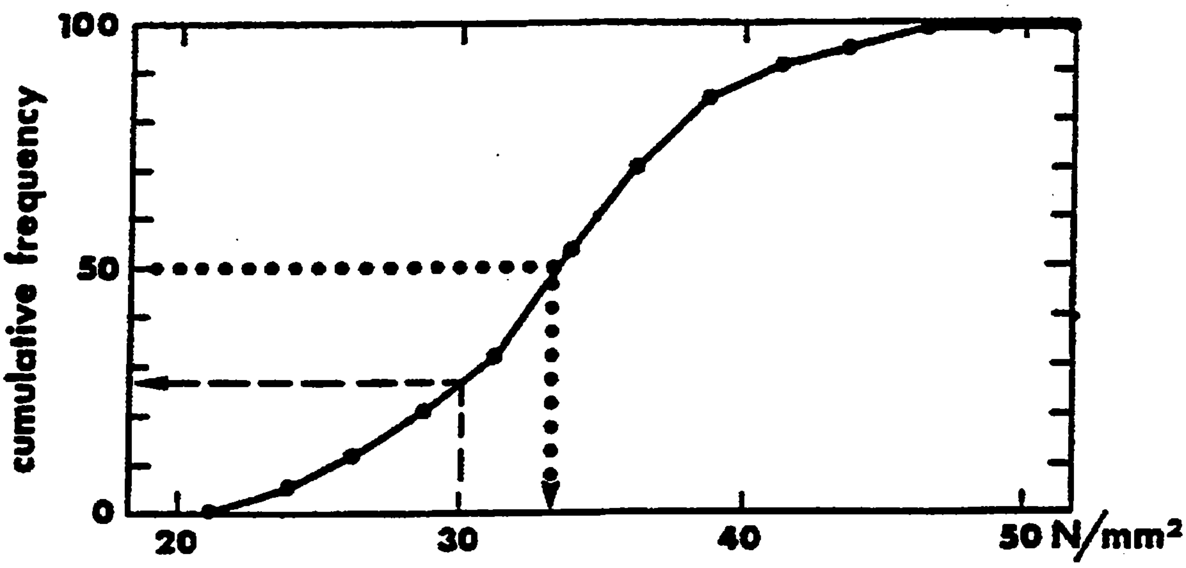
Relative (or absolute) frequencies plotted against class mid point. Every point on the curve corresponds to the relative (or absolute) frequency of observations falling in this class.



Cumulative frequency curve

Cumulated relative frequencies plotted against upper class boundaries. Every point on the curve corresponds to the portion of values which are smaller than any given strength x.

Example: Half of the measurements are smaller (respectively greater) than 33.5 N/mm^2



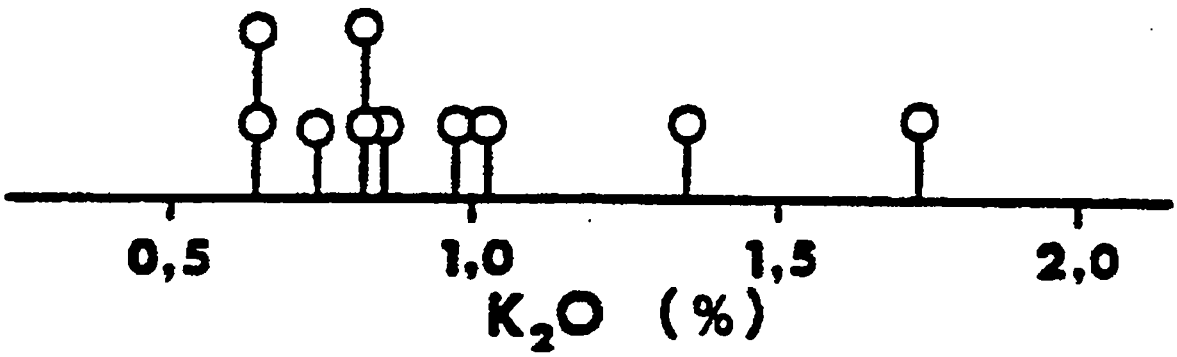
Stem-and-leaf Diagram

Similar to the histogram, but allows to see the individual data. Each observation is a leaf of a stem.

```
20   9
21
22   8
23   0579
24   456
25   0
26   1
27   126899
28   455
29H  039
30   0122378
31   456778
32   0244888
33M  033456
34   124569
35   2233455889
36H  14688
37   14689
38   169
39   247
40   12
41   09
42
43   4
44   3
45   38
     ***OUTSIDE VALUES***
48   9
```

Representation of small samples

For small sample sizes n the individual values are plotted directly on the measurement scale. Suitable to detect outliers and skewed distribution.

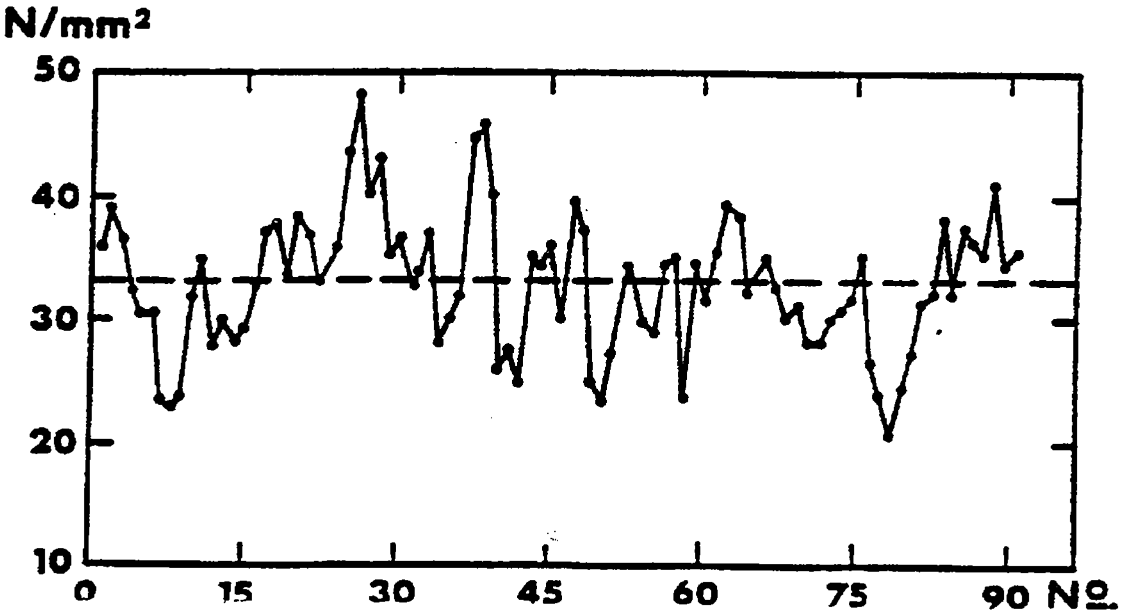


Example: K_2O-content of 10 cements

Time-plot

Often it is appropriate to draw the observations in chronological order to show a possible change of the level in time.

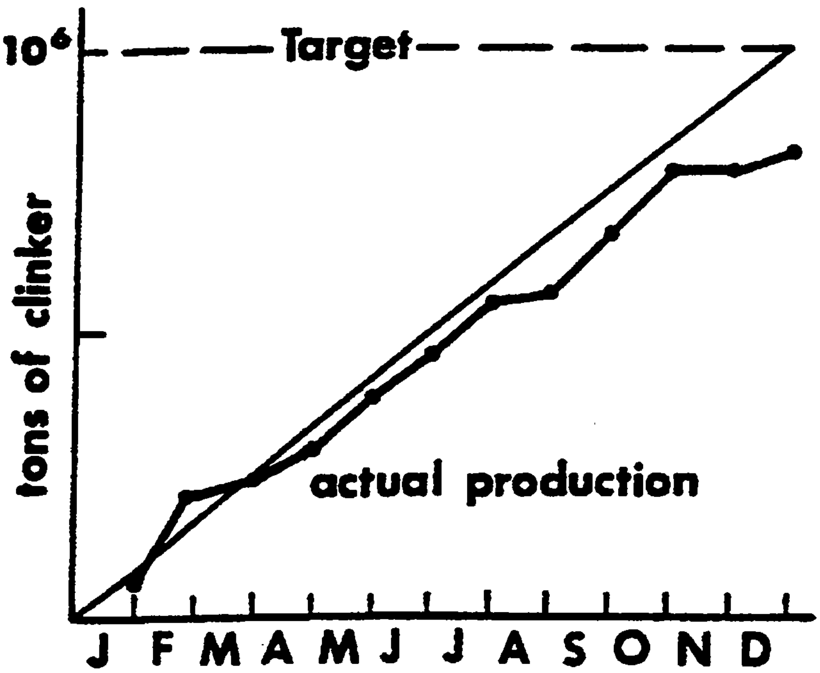
Useful in annual reports.



Cumulative time-plot

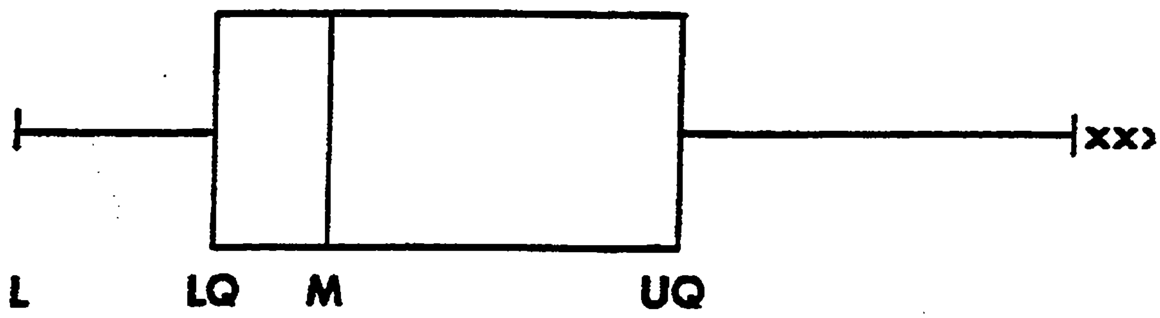
Cumulated values are plotted against time. Used to show deviations from a cumulative target.

Example: Actual clinker production is cumulated every month and compared with a target.



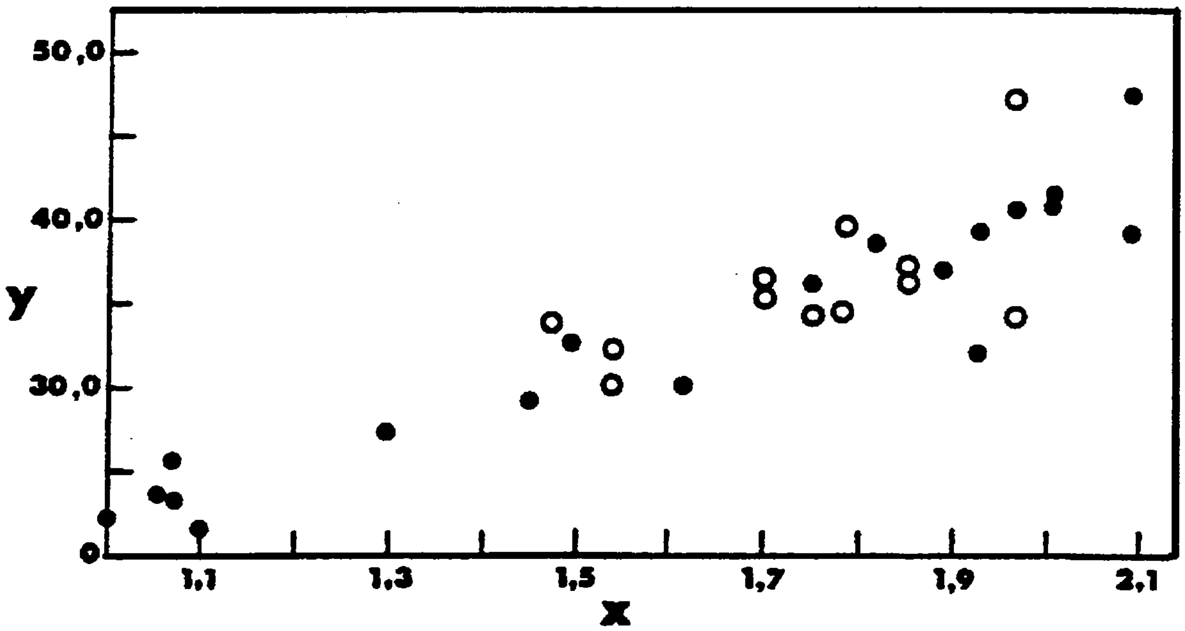
Box-Plot (Box-and Whisker-Plot)

Special and frequently used plot to numerize data. Shows median, lower and upper quartile (forming the box) and two lines (the whiskers) extending to the extremes of the data. If unusual extreme values occur, the whiskers extend only to those points that are within 1.5 times the inter-quartile range.



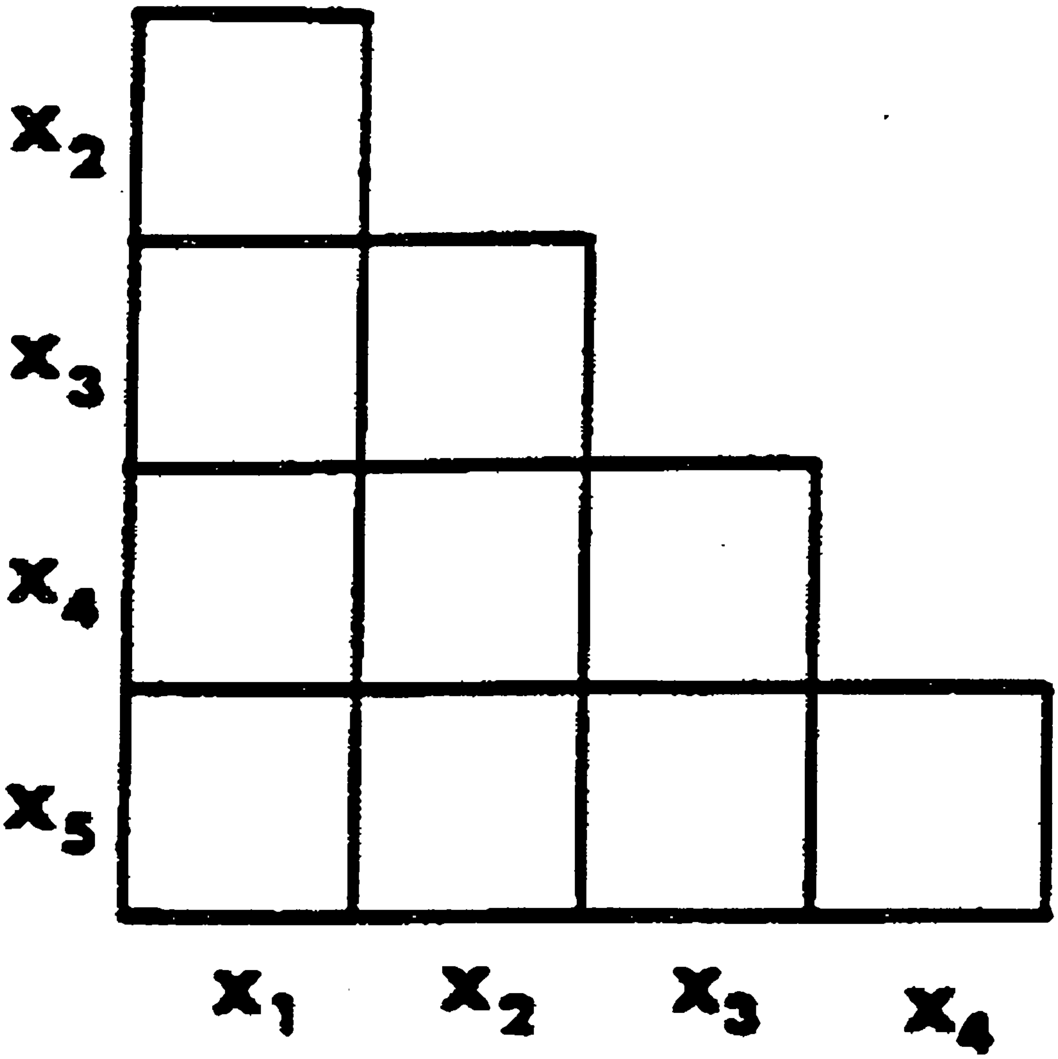
Scatter-plot or x-y plot

Plots two variables against each other.



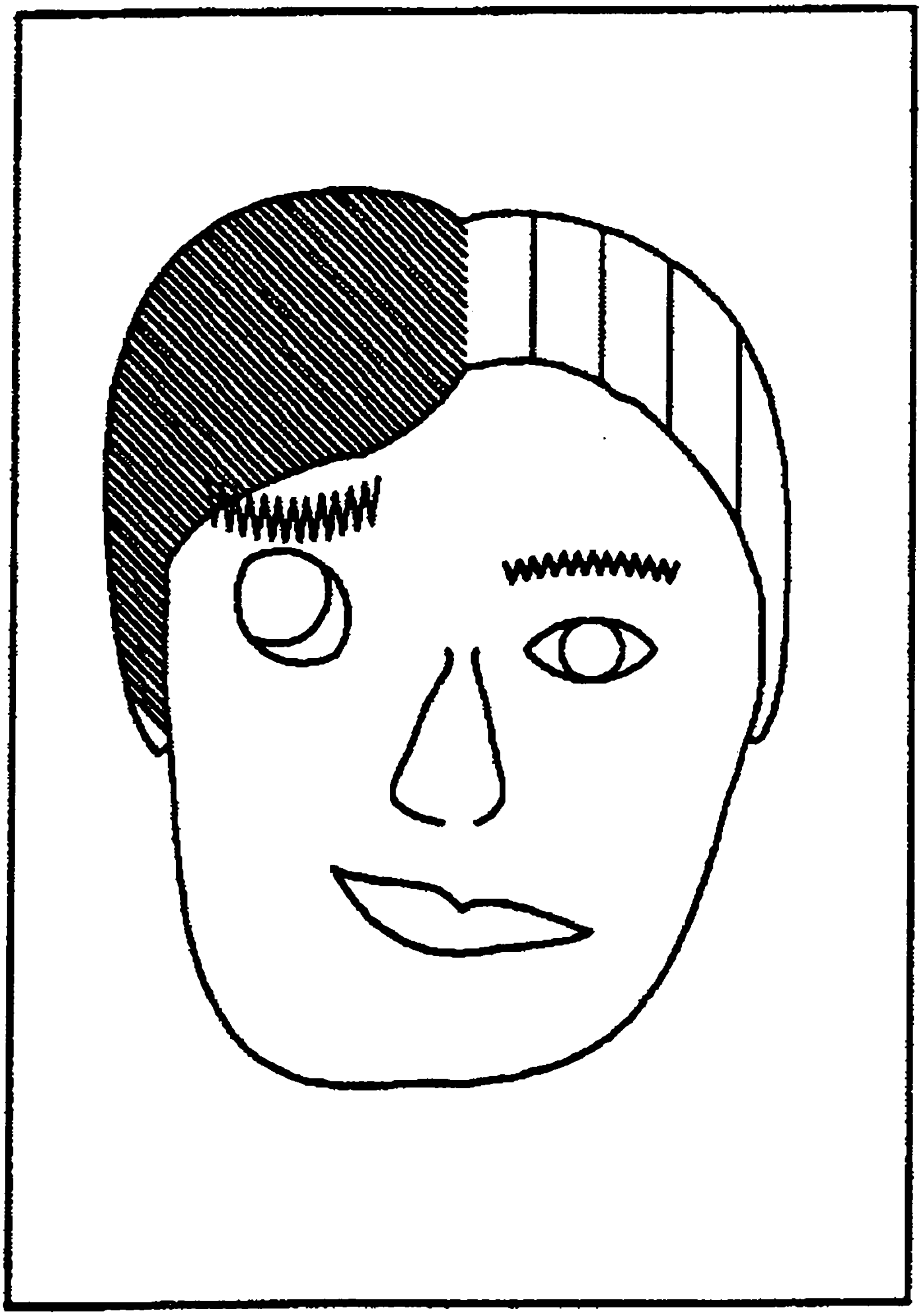
Draftmensplot

Plots two or more variables against each other!



Flury-Riedevyl faces and Chemoff faces

Faces are a method to show multivariate data graphically. Each variable is mapped to one or more of the face parameters. Other methods to represent multivariate data are the star symbol plot, the sun ray plot, castles and trees.



6.4 Control Charts

Control charts are special time-plots to show a possible change of the characteristics in a production process. The essence of a control chart is its clearness and intelligibility. In a quick survey it is possible to recognize natural subgroups, within which variation is likely to be random but among which assignable causes of one sort or another may cause non-random variation.

Summarized statistical errors (such as means or standard deviations) and confidence limits determined from overall sample data can be very misleading, if the data are not free from the effects of assignable causes. In cement industry annual means and standard deviations are mostly misleading due to a nonstationarity of the process, i.e. the occurrence of systematic changes at the production level. (Example: mean and standard deviation of the lime saturation of raw meal and/or clinker in annual reports. Changes in the target value required by the process or the product lead to a high non-interpretable standard deviation.)

If possible, the data should originally be collected with this subgrouping in mind; in such cases, the analyses can be simplified by arranging for an equal number of observations in each subgroup. If the data must be analyzed as they come, subgrouping may still be possible with knowledge of the data's source, i.e. obvious changes in the process conditions must be registered together with the data.

Control limits

To judge a significant change at the production level, the charts are completed with control limits. As a general rule it can be assumed that such a change is present, if a sample point falls outside of 3 σ-limits, where σ is the standard deviation of a homogeneous production phase.

This engineering rule has been found to work well. No exact probability is given for a chance variation beyond 3 σ-limits, but in general it is very s mall, in the order of perhaps 0.3%.

It may be of advantage to use two limits, such as a 2 σ-warning-limit and a 3 σ-action-limit. In this case, an action should already take place when two subsequent values lie between the two limits.

The use of 3 σ-limits bases on pure statistical considerations. It may occur that these limits are in conflict with the required tolerances of an external standard (e.g. process, standard or market requirements). If the required tolerances are smaller than the statistical ones, a modification of the process is necessary to improve its precision. In the opposite case, the statistical limits may be used in order to detect changes as early as possible.

Control Charts

Several types of charts may be used to detect different types of changes in the process:

	Control chart			
Cause of Change	**Mean X**	**Range R**	**Standard deviations**	**Cumulative**
Gross error (blunder)	1	2	-	3
Shift in average	2-	-	1	
Shift in variability	-	1	2	-
Slow fluctuation (trend)	2	-	-	1
Rapid fluctuation (cycle)	-1	2	-	

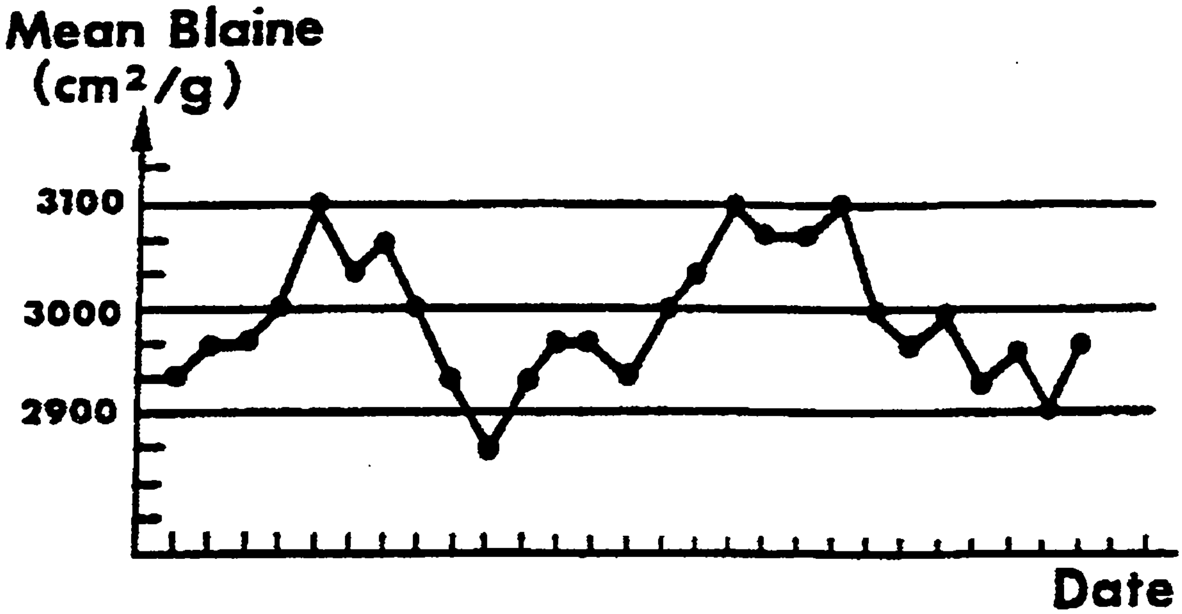
- = not appropriate / 3 = least useful

2 = useful / 1 = most useful

$\bar{x}$ – chart

Samples of fixed size n are taken from time to time. The arithmetic mean of the samples are plotted versus time.

Control limits (for warning and/or action) can be computed with the help of special procedures.



R-chart of s-chart

Given are small samples of fixed size n. The range or the standard deviation of the sample is plotted.

Values exceeding a certain control limit indicate an increase of variability.

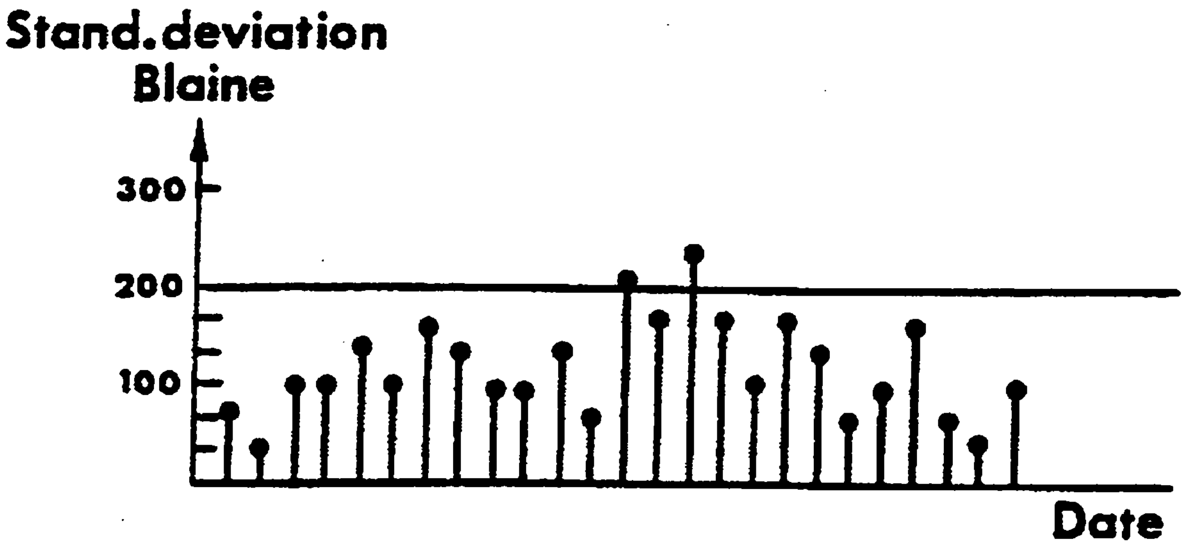
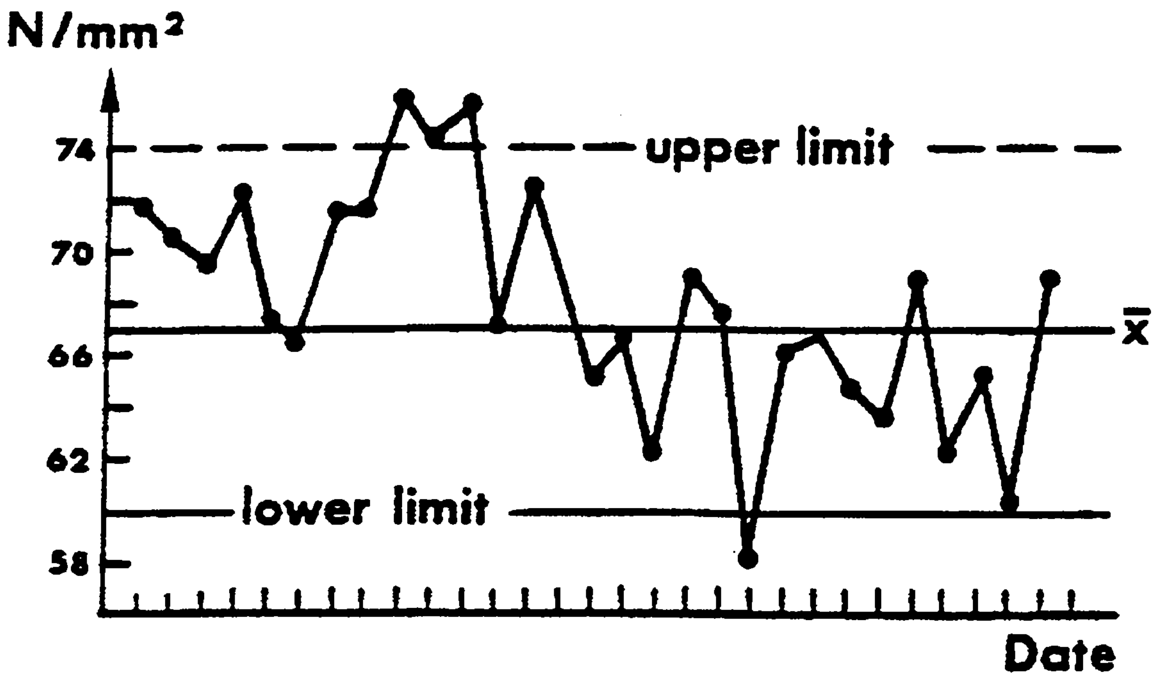


Chart of individual values

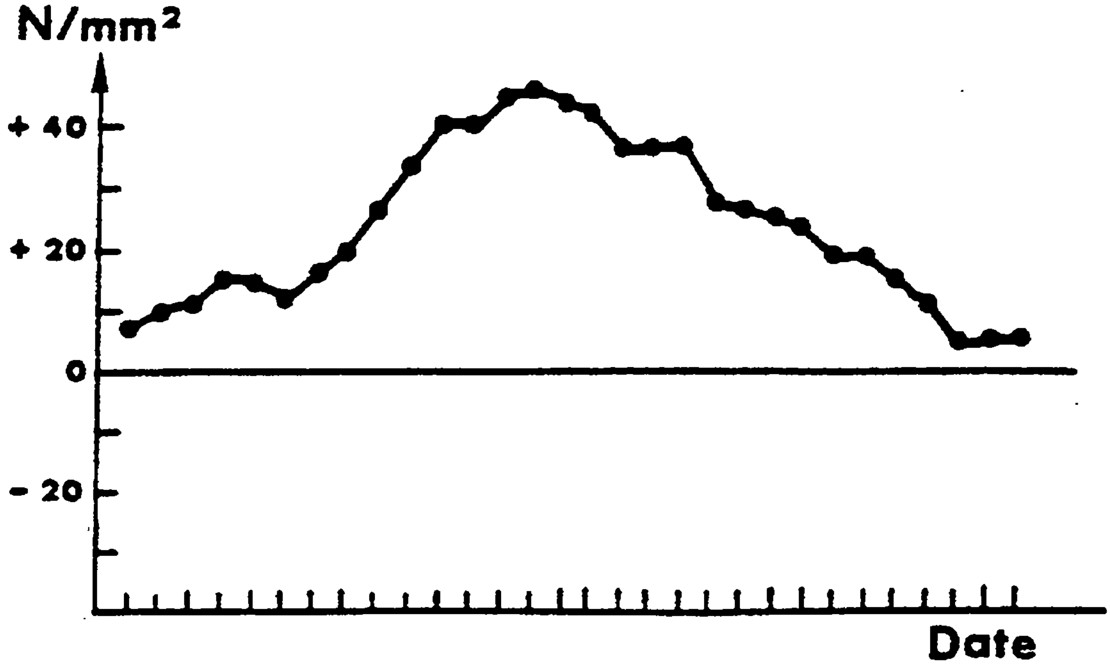
Instead of means, individual values ma y be plotted in the order of measurement.



CUSUM-chart

Cumulative-sum-chart. The deviations of individual values from a target value are cumulated.

This chart is especially sensitive to slow fluctuations in the process.

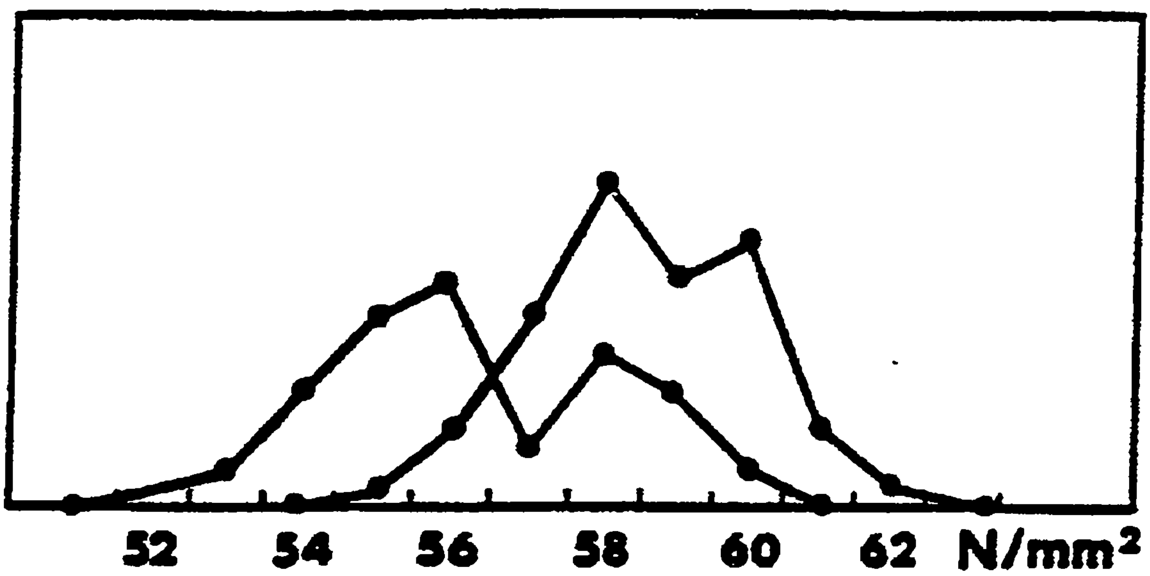


6.5 Comparative representation

Frequency polygon

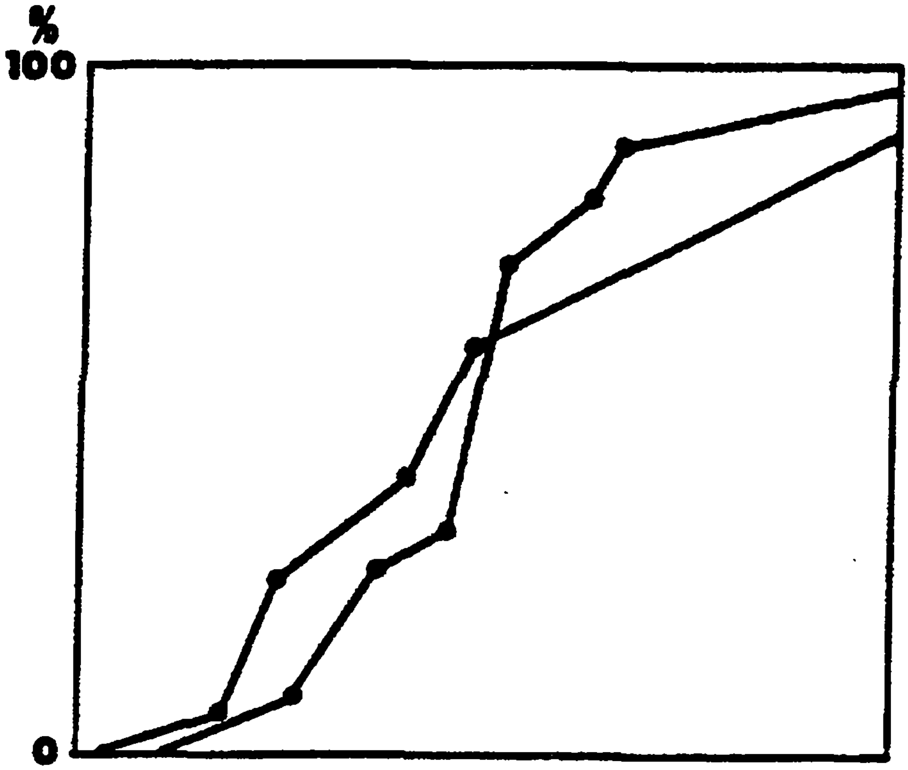
Comparison of the distribution of two samples. Relative frequencies corresponding to the histogram are plotted on a line graph against the mid-points of the classes.

Example: Mortar strength in the first and the second half-year.



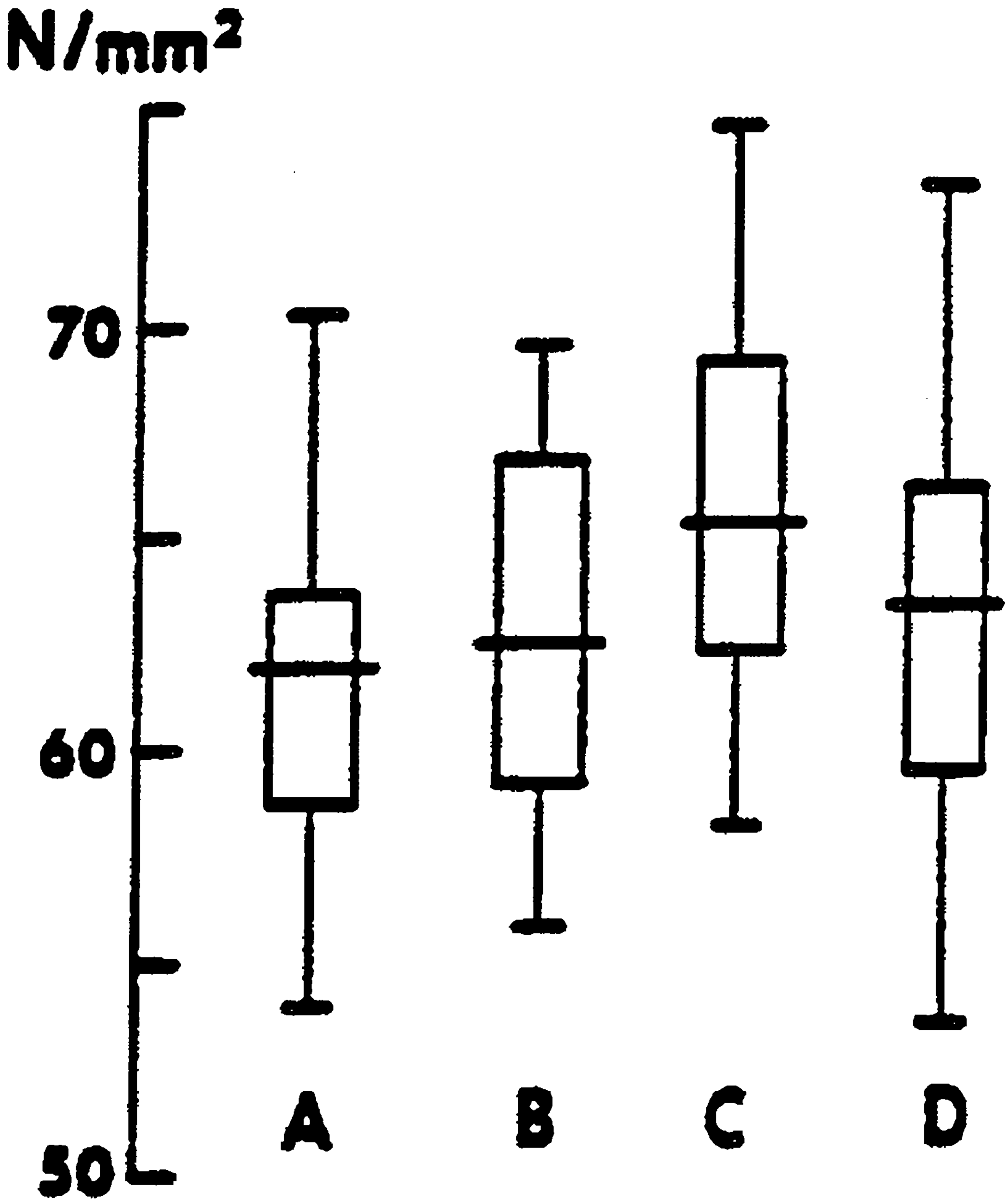
Cumulative Frequency plot

Comparison of the distribution of two samples. Relative cumulative frequencies are plotted against upper class limit.



Grouped Box-Plots

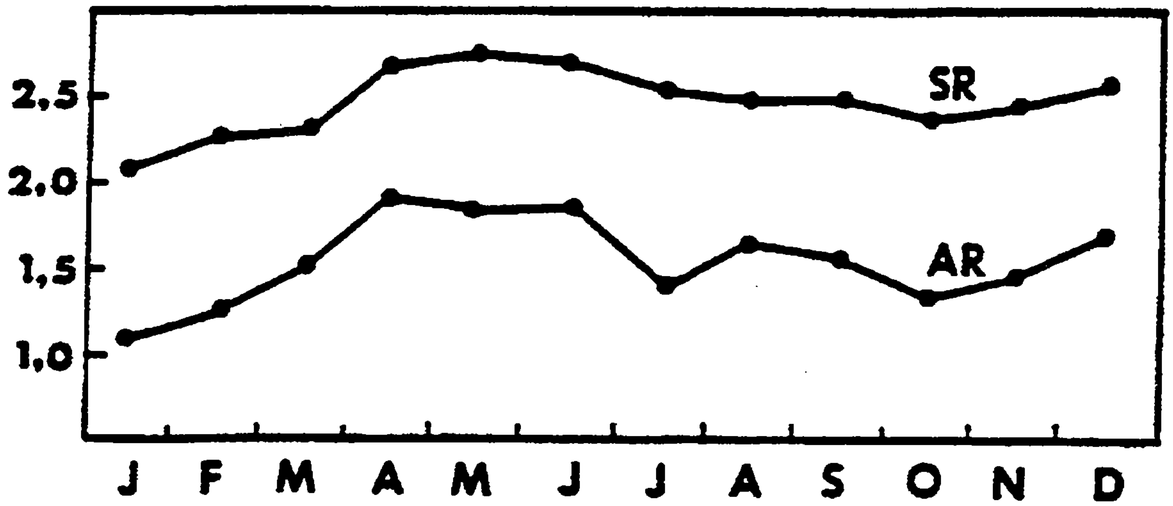
Shows two or more box-plots in the same graph. May be used to show the distribution of different variables or the distribution of the same variable in different groups (samples). The notched box plot may be used for the latter comparison.



Time-plot

Comparison of simultaneous changes of several variables in an observed period.

Example: Silica- and alumina ratio of a cement type in a specific year.

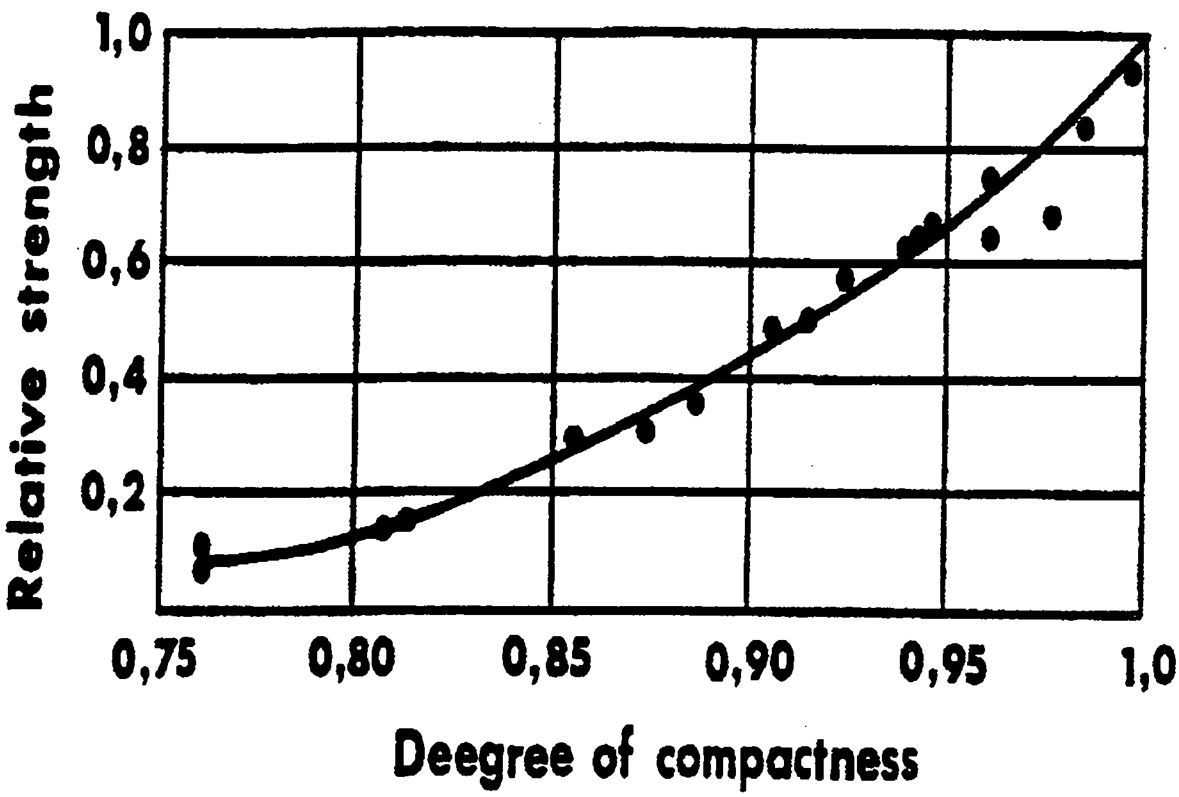


Scatter-plot

Illustration of the relationship between two variables.

Example: Relation between strength and density of concrete. The relationship is not linear.

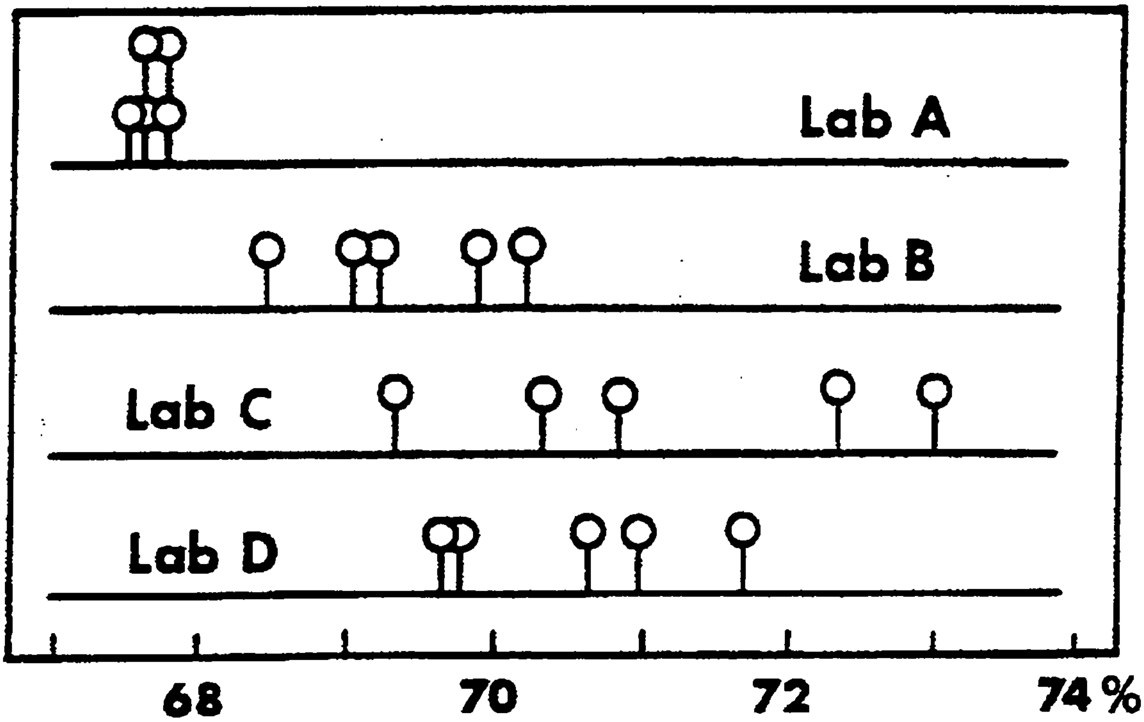
Don't use the correlation coefficient in this case, because it is only a measure for linear dependence.



Comparison of small samples

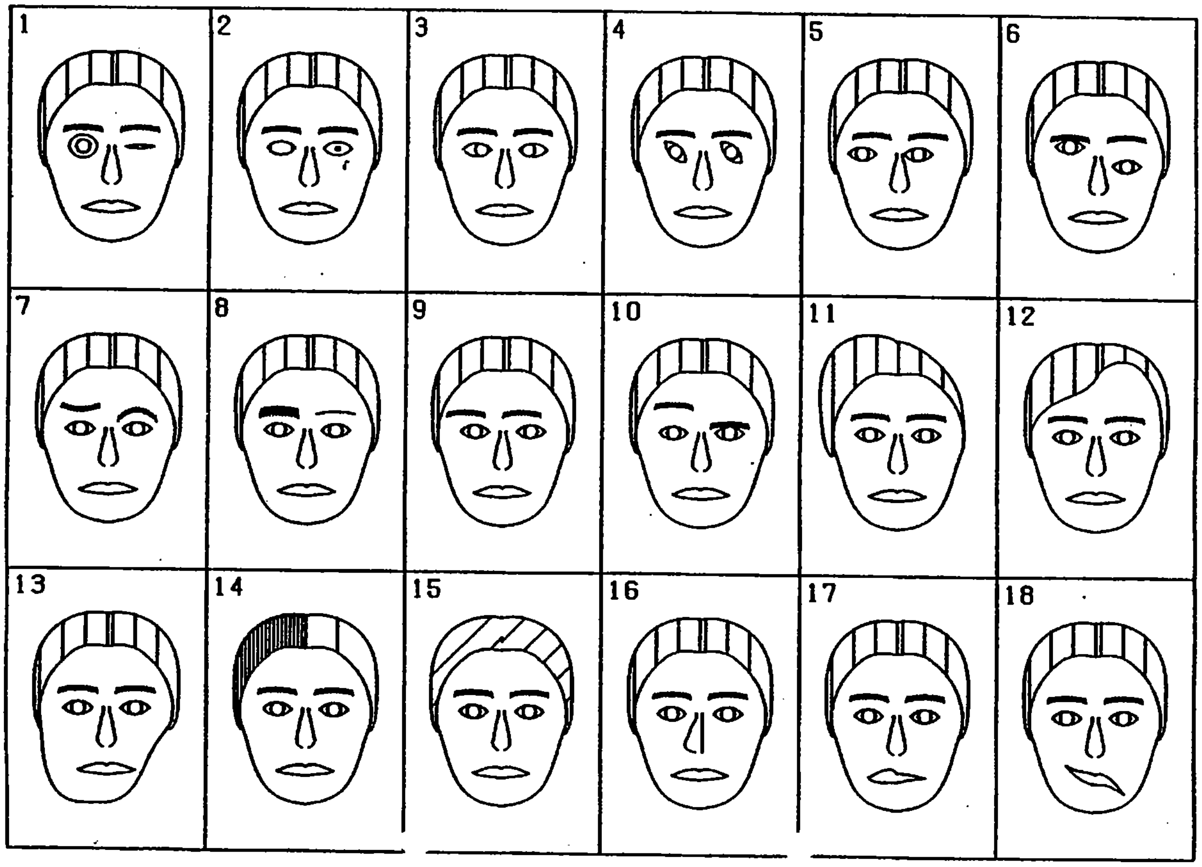
Individual values are marked on the scale separately for each group of observations.

Example: Inter-laboratory-test concerned with elite-content.



Asymmetrical Flury-Riedwyl Faces

This is a special version of the faces. E.g. the variables before and after treatment, respectively, are mapped to the same face parameter on the left and right face side, respectively. Or the left side shows the specification values of the variables and the right side the actual measurements.



7. CORRELATION AND REGRESSION

7.1 Correlation coefficient

The degree of linear dependence between two random variables X and Y can be expressed by the correlation coefficient r_{xy}.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

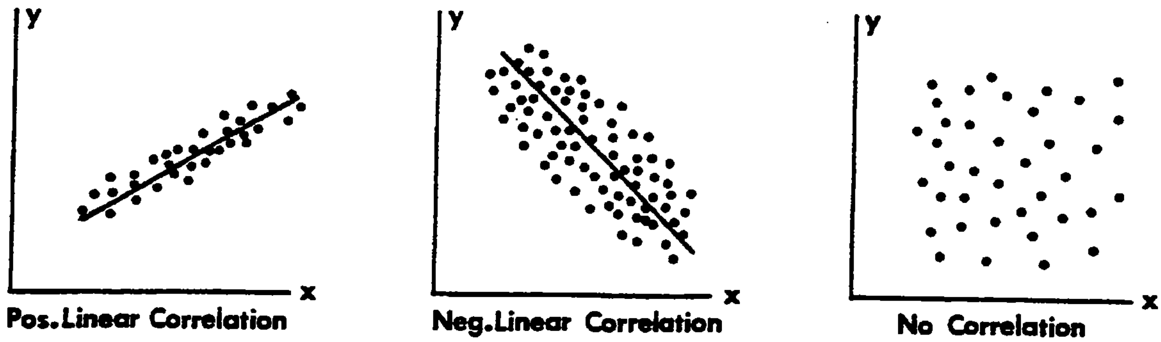
with $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ the covariance of X and Y

and s_x, s_y standard deviations of X and Y.

For practical computations use

$$s_{xy} = \frac{1}{n-1}\sum x_i y_i \frac{1}{n}(\sum x_i)(\sum y_i)$$

The following figure shows the scatter diagrams for several degrees of dependence.



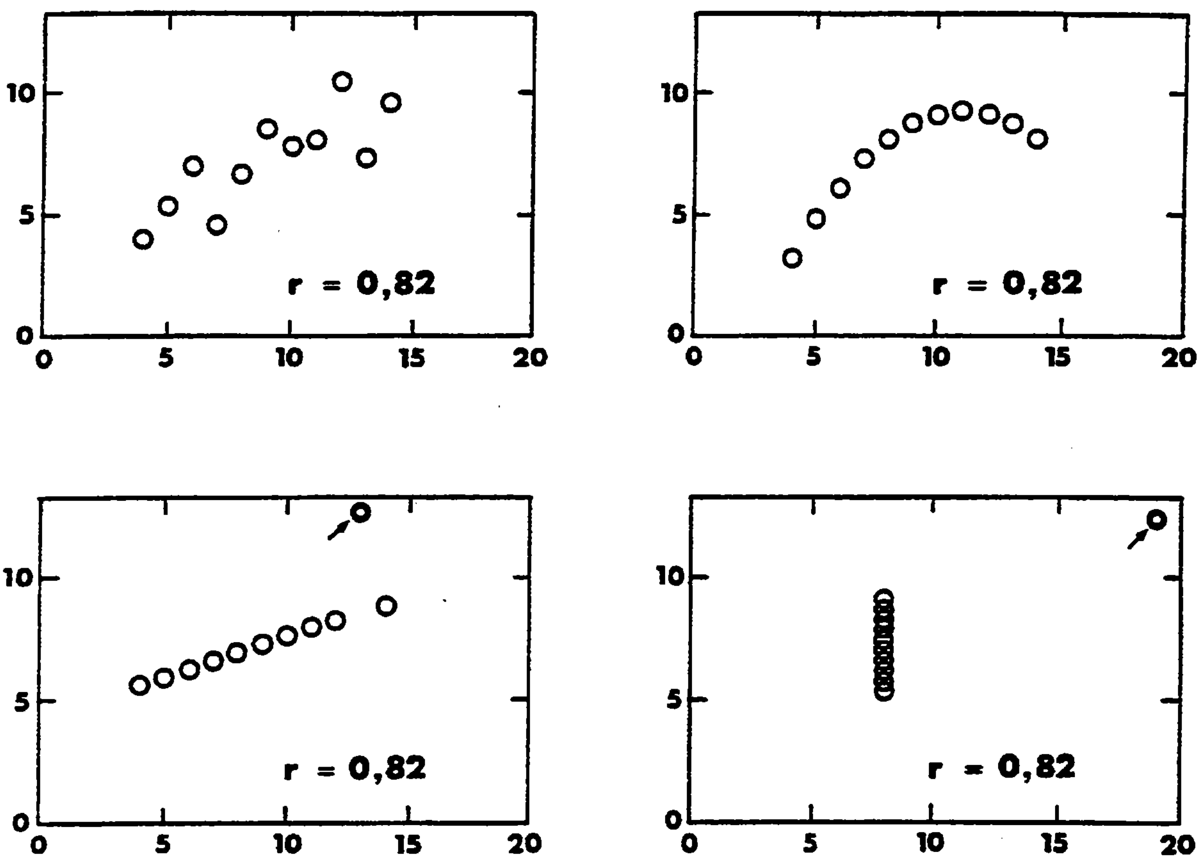
Properties of r_{xy}:

a) r_{xy} = ranges from -1 to +1
r_{xy} = +1 : All measured values lie on an increasing line
-1 : All measured values lie on a decreasing line
0 : No linear relationship between the measured values

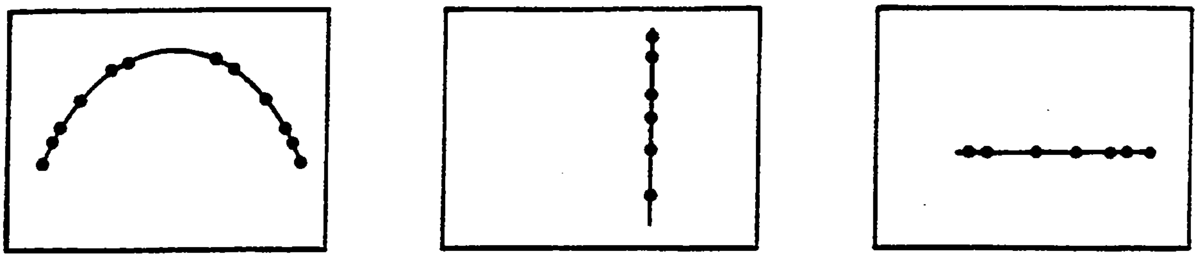
b) r_{xy} is a measure of linear dependence. If the scatter diagram indicates a non-linear relationship, then the correlation coefficient will be misleading and should not be calculated.

c) r_{xy} is very sensitive (not robust) against outliers.

The necessity of drawing a scatter diagram is illustrated in the following figures. Completely different graphs may result with equal correlation coefficients (r = 0.82).



r_{xy} can be close to zero even though the variables are clearly non-linear dependent and r_{xy} is not defined if s_x or s_y is zero.



r = 0 **r is not defined**

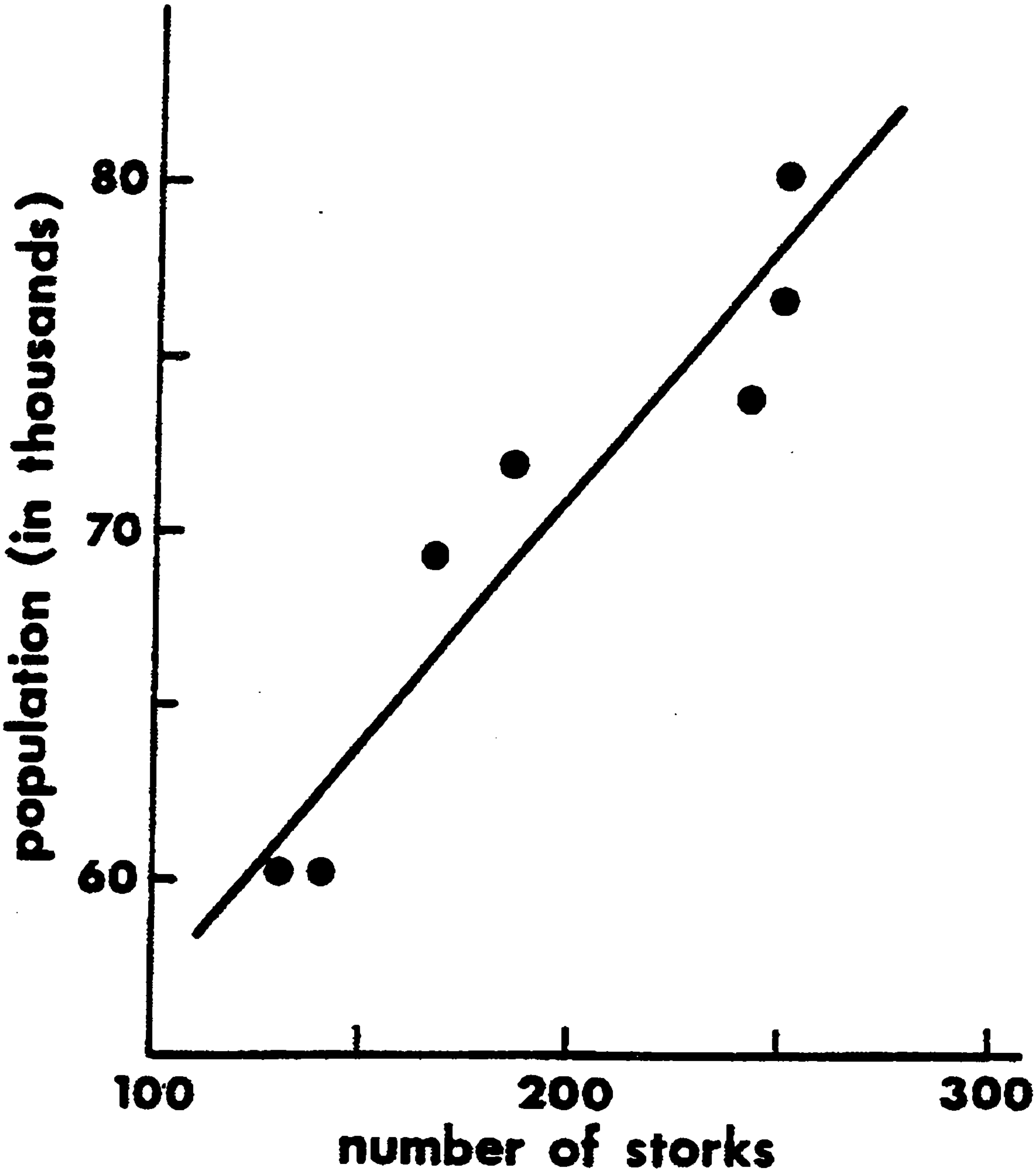
Interpretation of r_{xy}

A high correlation coefficient between two variables does not necessarily indicate a causal dependence. There may be a third variable not under control which is causing the simultaneous change in the first two variables, and which produces a spuriously high correlation coefficient. In order to establish a causal relationship it is necessary to run a carefully controlled experiment (see chapter 8). Unfortunately it is often impossible to control all the variables which could possibly be relevant to a particular experiment, so that the experimenter should always be on the lookout for spurious correlation.

The following is an example for confusion of correlation with causation:

The following figure shows the population of Oldenburg at the end of each of 7 years plotted against the number of storks observed in the corresponding year. Although in this example few would be led to hypothesize that the increased number of storks caused the observed increase of population, investigators are sometimes guilty of this kind of mistake in other contexts. Correlation between two variables Y and X often occurs because they are both associated with a third factor W. In the stork example, since the human population Y and the number of storks X both increased with time W over this 7-year period, it is readily understandable that a correlation appears when they are plotted together as Y versus X.

Fig.: **A plot of the population of Oldenburg at the end of each year against the number of storks observed in that year.**



By using sound principles of experimental design and, in particular, randomization, data can be generated that provide a more sound basis for deducing causality.

7.2 Linear Regression

If the scatter diagram indicates that the two variables are linearly related, then we may want to predict the value of one of the variables from a given value of the other variable. For this purpose a regression line is fitted to the data (method of least squares).

Model assumptions

$$y_j = \alpha + \beta_j + E_j \qquad (1)$$

where the residuals E_j's are independent normal random variables with mean 0 and constant variance σ2.

α = y -intercept

β = slope

x = independent variable - Y dependent variable

This assumption is essential if parameter reductions (see later), i.e. reduced models are tested and the assumptions have to be checked when a model is fitted.

The estimates of the parameters α and β in model (1) are obtained by the method of Least Squares (MLS), i.e. the estimates are obtained by minimizing the sum of squared residuals

$$\sum_{j=1}^{n}(y_j - \alpha - \beta x_j)^2 = \sum_{j=1}^{n}\hat{E}_j^{\,2}$$

with respect to α and β !

This method is very wide spread in statistics and the estimates have good statistical properties under the above model assumptions! Today there are also procedures which do not minimize the sum of squared residuals but use other criteria (Robust Methods!).

In the following let

$$s_{xx} = \sum(x_j - \bar{x})^2 = \sum x_j^{\,2} - \frac{1}{n}(\sum x_j)^2 = (n-1)s_x^{\,2}$$

$$s_{yy} = \sum(y_j - \bar{y})^2 = \sum y_j^{\,2} - \frac{1}{n}(\sum y_j)^2 = (n-1)s_y^{\,2}$$

$$s_{xy} = \sum(x_j - \bar{x})(y_j - \bar{y}) = \sum x_j y_j - \frac{1}{n}\sum x_j y_j = (n-1)s_{xy}$$

The resulting estimated regression line is

$$\hat{y} = \hat{\alpha} + \hat{\beta}\text{, where } \hat{\beta} = \frac{s_{xy}}{s_{xx}}$$

$$\hat{\alpha} = \bar{y} + \hat{\beta}\bar{x}$$

with a resulting minimal sum of squares (MSSQ) $s_{\min} = s_{yy} - \frac{s^2{}_{xy}}{s_{xx}}$ with n-2 degrees of freedom (2 parameters α and β are estimated). $\hat{y}$ is the predicted value for the dependent variable y.

Note:

1) In the mentioned problem there are two regression lines, one to predict y from x and one to predict x from y. The two lines are not identical and therefore it is not allowed to invert the regression equation.

2) The prediction equation is valid only in the observed range of observations. Extrapolations may give misleading results. We have no information that linearity holds outside of the present observations.

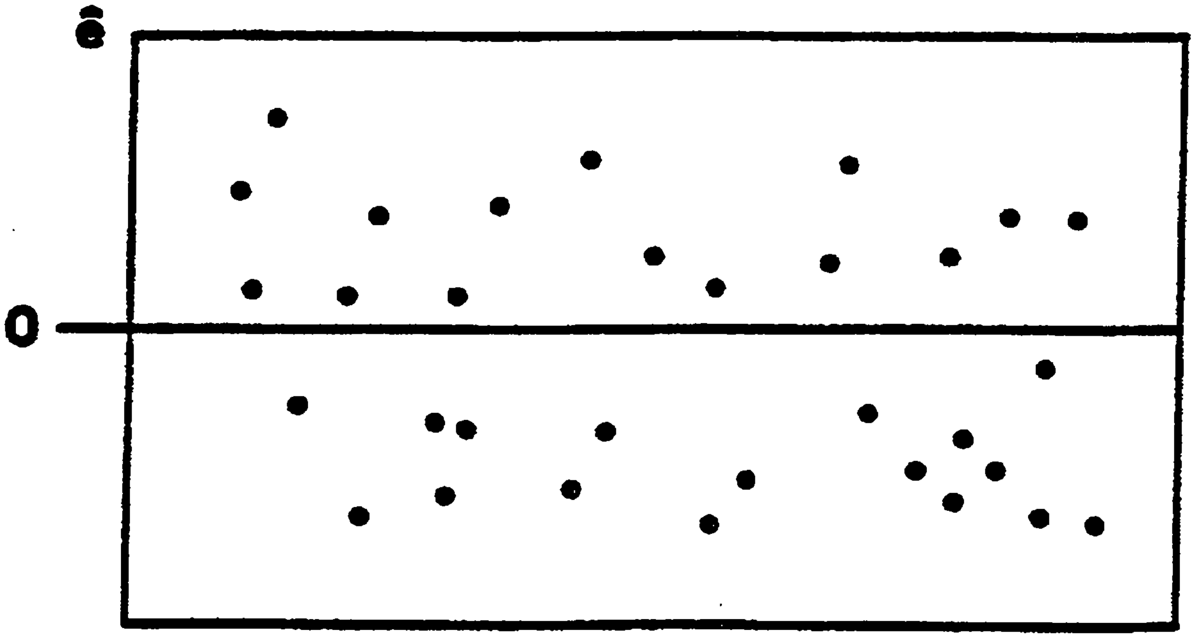
For each pair (x_j, y_j) we can therefore compute the predicted value $\hat{y}$ of y by means of the regression function $\hat{y} = \hat{\alpha} + \hat{\beta}x$ i.e.

$$\hat{y}_j = \hat{\alpha} + \hat{\beta}x_j$$

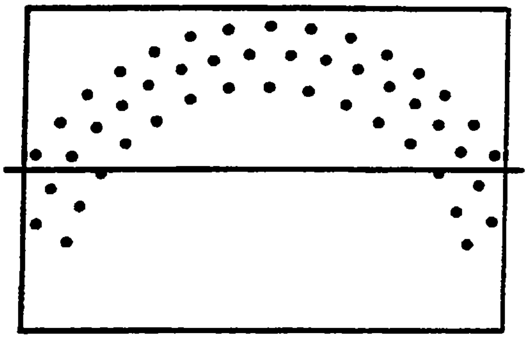
and the estimated residual

$$\hat{e}_j = y - \hat{y}_j$$

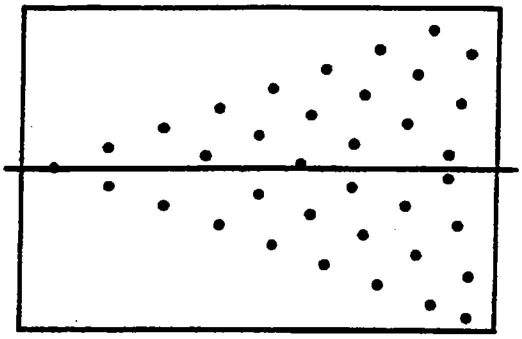
$\hat{\alpha}^2 = S_{min}/(n-2)$ is an estimate for the variance of the residuals. The analysis of the residuals $\hat{e}_j$ gives us the possibility to validate the model assumptions (Independence, Normality and constant Variability). The first check is done by plotting the residuals against the predicted y_j. The residuals $\hat{e}_j$ should be randomly scattered around the line e = 0 and show no pattern.



For example the following structures of the residuals would indicate, that the model is not correctly specified (e.g. that the variables x and y should be transformed before calculation of the regression line).



Relation not linear
(e.g. $y = h(\alpha + \beta x)$
or $y = \dfrac{1}{\alpha + \beta x}$



Variance not constant
(e.g. variance σ^2 proportional to x)

The second check would be a normal probability plot of the residuals.

If the residuals suggest that the assumptions do not hold, then further investigations are necessary (suitable transformations, other influencing variables, time-dependence => growth curves, time-series-analysis).

After confirmation of the assumption and calculation of the regression line (sometimes also before calculation) it is possible that we want to investigate whether perhaps the y-intercept equals zero, or the slope equals zero or whether there is no relation between y and x ($\alpha = \beta = 0$) or whether the α or β equal some specified values α_0 or β_0 (e.g. if a value of $\hat{\beta} = 1.1$ may be replaced by $\beta_0 = 1.0$). This means that we want to test if a reduced (simplified) model is sufficient to describe the relationship between y and x. For this it is essential that the model assumptions hold.

7.2.1 Regression line with slope 0:

Model: $y_j = \alpha + E_j$

Estimate $\hat{\alpha}$ for α : $\hat{\alpha} = \bar{y}$ (= mean of the y_j's)

Minimal Sum of Squares (MSQ): $S_{min} = S_{yy}$

Degrees of freedom (df): df = n-1

7.2.2 Regression line with y-intercept 0

Model: $y_j = \beta x_j + E_j$

Estimate $\hat{\beta}$ for β: $\hat{\beta} = \dfrac{\sum x_j y_{j1}}{\sum x_j^2}$

MSQ: $S_{min} = \sum y_j^2 - \frac{(\sum x_j y_j)^2}{\sum x_j^2}$.

df = n-1

7.2.3 Intercept = 0. Slope = 0

Model: $y_j = E_j$

no parameters to estimate

$$MSQ = \sum y_j^2$$

df = n

Other models:

$$y = \alpha_0 + \beta x$$

$$y = \alpha + \beta_0 x$$

$$y = \alpha_0 + \beta_0 x$$

The estimates and MSQ can be obtained by differentiation of the corresponding Sum of Squared Residuals!

These special models are only useful if the data really speaks for them!

7.2.4 Comparison of models

This is done by performing an Analysis of Variance (ANOVA). ANOVA is a widely spread statistical technique. (Comparison of more than two means, Experimental Designs etc.). ANOVA compares the Minimal Sum of Squares of the models to each other. The basic results for the comparison are filled in a table, the ANOVA-Table.

ANOVA-TABLE

Model	MSQ	df
H_0	S^0_{min}	df^0
A	S_{min}	df
Reduction	$S^0_{min} - S_{min}$	df_0 - df

H_o denotes the null hypothesis, A the alternative model. By H_0 the test-statistic

$$F = \frac{(S^0_{min} - S_{min})/(df^0 - df)}{S_{min}/df}$$

is distributed according to a F-distribution with $m_1 = (df^0 - df)$ and $m_2 = df$ degrees of freedom.

Significance limits: Look up $F_{1-\alpha;(df^0-df),df}$ in Table A-9 with $m_1 = (df^0\text{-}df)$ and $m_2 = df$ degrees of freedom.

Decision: If $F > F_{1-\alpha;(df^0-df),df}$ then conclude that the simplification to the null-model H_0 is not permitted, therefore the alternative model has to be used; otherwise ($F \leq F_{1-\alpha}$) there is no evidence that the null-model should be rejected.

7.2.5 Standard deviation of the estimates

The ANOVA of the null hypotheses $H'_0 : y = \alpha$ and $H''_0 : y = \beta x$, resp. against the alternative $A : y = \alpha + \beta x$ allows us to compute the standard deviations (sd) of the estimated parameters $\hat{\alpha}$ and $\hat{\beta}$ in the alternative model $A : y = \alpha + \beta x$: let $F(\beta = 0)$ be the calculated F-statistic for the test H'_0 against A (e.g. ANOVA for $\beta = 0$) and correspondingly $F(\alpha = 0)$ for H''_0 against A.

Then

$$sd(\hat{\alpha}) = \frac{|\hat{\alpha}|}{\sqrt{F(\alpha = 0)}}$$

$$sd(\hat{\beta}) = \frac{|\hat{\beta}|}{\sqrt{F(\beta = 0)}}$$

so that the regression line of the alternative (full) model is

$$y = \underset{(sd(\hat{\alpha}))}{\hat{\alpha}} + \underset{(sd(\hat{\beta}))}{\hat{\beta}} x$$

In order to obtain the standard deviation in a model $A : y = \beta x$, this model has to be compared with the null-model $H_0 : y = 0$.

For $A : y = \alpha$, compare A to $H_0 : y = 0$

For $A : y = \alpha_0 + \beta_x$, compare A to $H_0 : y = \alpha_0$

For $A : y = \alpha + \beta_0 x$, compare A to $H_0 : y = \beta_0 x$

$1-\alpha$ confidence intervals for the estimated parameters can be obtained by calculating

$$\hat{\alpha} \pm t_{1-\alpha/2,m} sd(\hat{\alpha})$$

and

$$\hat{\beta} \pm t_{1-\alpha/2,m} sd(\hat{\beta})$$

where m equals the degrees of freedom in the alternative model and $t_{1\_a/2,m}$ can be looked up in Table A-4.

7.2.6 Coefficient of determination

The prediction of y with a regression line is more or less accurate dependent on the degree of linear dependence. How well does the regression line fit the data? A measure to express the relative accuracy of prediction compared with the total variation of y is the coefficient of determination r^2. In the case of a regression line the coefficient of determination is the square of the correlation coefficient

$$r^2 = r_{xy}^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

The total variation of y can be partitioned into two components. Total variation = explained variation + unexplained variation

where

total variation	$= \sum (y_i - y)^2$	
unexplained variation	$= \sum (y_i - \hat{y})^2$	(observed - predicted)
explained variation	$= \sum (\hat{y} - \overline{y})^2$	

The coefficient of determination is the ratio

$$r^2 = \frac{\text{explainedvariation}}{\text{totalvariation}}$$

In this form r^2 is defined also for non-linear curves and multiple regression.

7.2.7 Transformations before Regression Analysis

Kind of transformation	Type of relation	Regression after transformation
$y' = \ln y, x (y > 0)$	$y = \alpha e^{\beta x}$ (exponential)	$y' = \ln + \beta x$
$y, x' = \ln x, (x > 0)$	$y = \alpha + \beta \ln x$ (logarithmic)	$y = \alpha + \beta x'$
$y' = \ln y, x' = \ln x, (x, y > 0)$	$y = \alpha x^{\beta}$ (exponentiation)	$y' = \ln \alpha + \beta x'$
$y' = 1/y, x (y \neq 0)$	$y = (\alpha + \beta x)^{-1}$	$y' = \alpha + \beta x$
$y, x' = 1/x (x \neq 0)$	$y = \alpha + \beta \frac{1}{x}$	$y = \alpha + \beta x'$
$y' = 1/y, x' = 1/x (x, y \neq 0)$	$y = \frac{x}{\alpha x + \beta}$ (hyperbolic)	$y' = \alpha + \beta x'$
$y' = \ln y, x' = 1/x (y > 0, x \neq 0)$	$y = \alpha e^{\beta / x}$	$y' = \ln \alpha + \beta x'$
$y' = 1/y, x' = e^{-x} (y \neq 0)$	$y = \frac{1}{\alpha + \beta e^{-x}}$	$y' = \alpha + \beta x'$

Model Overview for Linear Regression

Model Overview for Linear Repression

Model	Parameter Estimate(s)	MSQ (=Smin)	dt
$\alpha + \beta x$	$\beta = S_{xy}/S_{xx}$ $\alpha = y - \beta x$	$S_{yy} - \left(\frac{S_x^2 y}{S_{xx}}\right)$	n - 2
$\alpha_o + \beta x$ (α_0 fest)	$\beta = \frac{\Sigma x_i y_i - \alpha_o \Sigma x_i}{\Sigma x^2}$	$\Sigma y_i^2 + n\alpha_o [\alpha_o - 2\bar{y}] - \frac{(\Sigma x_i y_i - \alpha_o \Sigma x_i)^2}{\Sigma x_i^2}$	n - 1
βx ($\alpha_o = 0$)	$\beta = (\Sigma x_i y_i)/\Sigma x^2_i$	$\Sigma y_i^2 - (\Sigma x_i y_i)^2/\Sigma x_i^2$	n - 1
$\alpha + \beta_o x$ ($\beta_o$fest)	$\alpha = y - \beta_o x$	$S_{yy} - 2 \beta_o S_{xy} + \beta_o^2 S_{xx}$	n - 1
α ($\beta_o = 0$)	$\alpha = y$	S_{yy}	n - 1
$\alpha + x$ ($\beta_o = 1$)	$\alpha = y - x$	$S_{yy} - 2 S_{xy} + S_{xx}$	n - 1
$\alpha_o + \beta_o x$	---	$S_{yy} + n[\alpha_o - (y - \beta_o x)]^2$ $+ \beta^2_o S_{xx} - 2 \beta_o S_{xy}$	n

7.2.8 Multiple and non-linear regression

Often the variable to be predicted is not only dependent on one but on several independent variables.

Model:

$$Y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_p x_{pj} + E_j$$

where E_j are independent normal random variables with mean 0 and constant variance σ^2. Again these assumptions have to be checked after a model has been fitted.

In this case prediction can be improved in a multiple (or multivariate) regression expressed as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \cdots + \hat{\beta}_p x_p$$

Or the scatter diagram shows a non-linear relationship which demands to fit a non-linear curve to the data. Often a logarithmic or some other transformation of one or both variables may lead to a linear relationship, and a regression line can be fitted to the transformed data. In more complex situations further methods are available.

For multiple, non-linear and mixed (multiple/non-linear) procedures consult literature (ref. Chatfield 1975)

Flury + Riedwyl (1988) give a very good praxis oriented introduction to multiple linear regression and multivariate analysis, including Discriminant Analysis, Principal Components, Identification and Specification analysis.

Example 6

The following data represent

y: water requirement (%) and

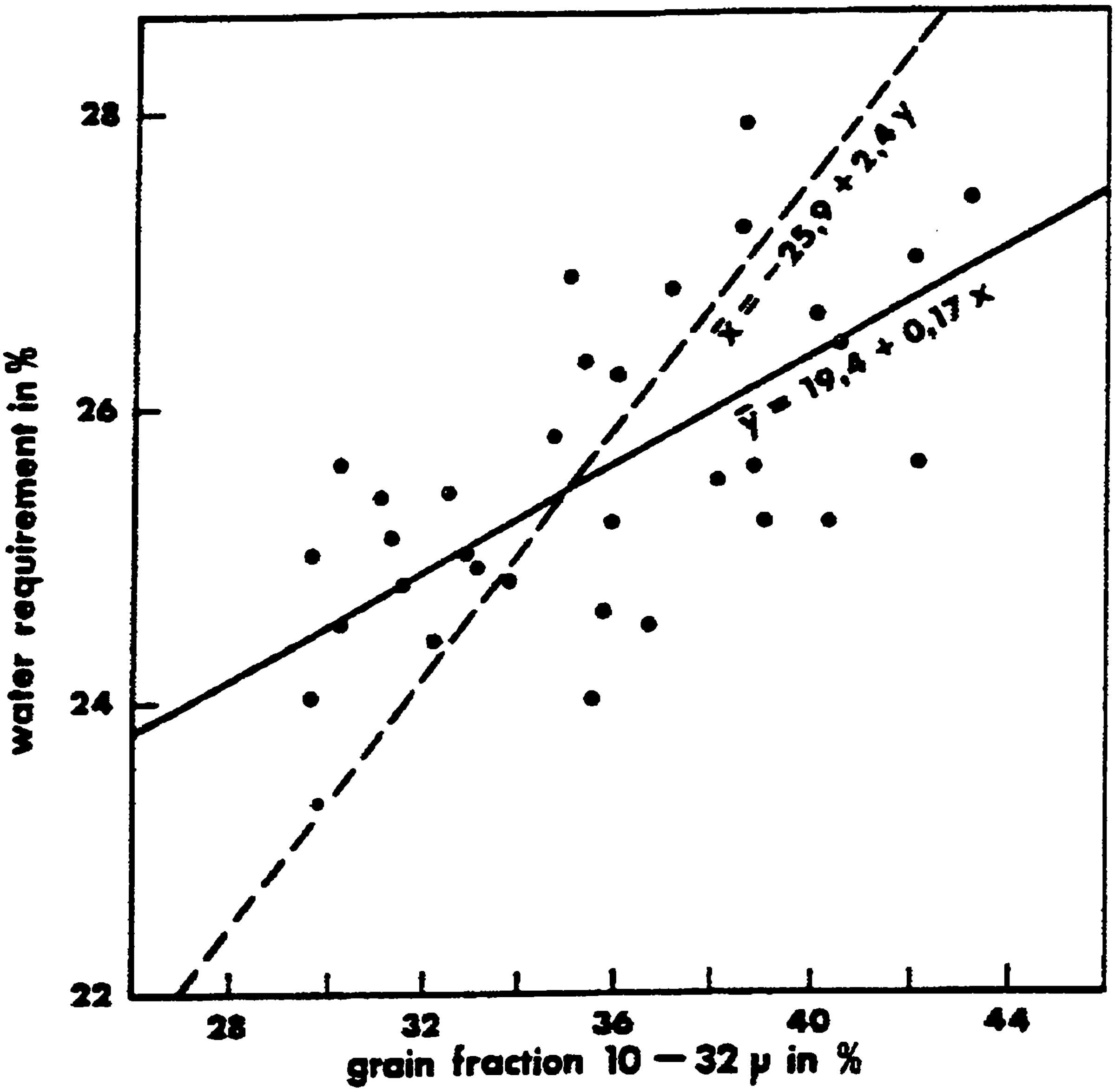
x: grain fraction 10-32 μ (%)

of n = 34 different cements.

x	y	x	y	x	y	x	y
32.9	25.0	38.6	27.2	29.8	23.3	37.2	26.8
35.4	26.3	32.5	25.4	30.3	25.6	42.2	25.6
33.1	24.9	31.3	25.1	36.0	25.2	35.8	24.6
38.7	27.9	35.1	26.9	30.2	24.5	39.1	25.2
33.8	24.8	32.2	24.4	35.6	24.0		
32.9	25.0	31.5	24.8	40.6	26.4	40.1	26.6
29.6	24.0	31.1	25.4	40.3	25.2	43.2	27.4
29.7	25.0	38.8	25.6	38.1	25.5	36.7	24.5
36.1	26.2	42.1	27.0	34.7	25.8		

The scatter diagram shows a positive linear relationship between water requirement and grain size. Regression lines to predict x or y are clearly distinct. The lines are identical if r_{xy} is 1 or -1.

If r = 0 the lines are perpendicular, i.e. the best prediction of y is the arithmetical mean $\bar{y}$. In the same way $\bar{x}$ is the best prediction of x if no dependence exists.



Computations

In a first step five auxiliary sums are provided:

$A = \Sigma x_i$ = 1'205.3 n = 34

$B = \Sigma x_i^2$ = 43'253.19

$C = \Sigma y_j$ = 867.1

$D = \Sigma y_i^2$ = 22'150.93

$E = \Sigma x_j y_j$ = 30'828.44

$$s_{xy} = \frac{1}{n-1}\left[E - \frac{AC}{n}\right] = 2.719545$$

$$s_x^2 = \frac{1}{n-1}\left[B - \frac{A^2}{n}\right] = 15.918333$$

$$s_y^2 = \frac{1}{n-1}\left[D - \frac{C^2}{n}\right] = 1.131203$$

$$\bar{x} = A/n = 35.45$$

$$\bar{y} = C/n = 25.05$$

$S_{xy} = 89.7550$

$S_{xx} = 525.3050$

$S_{yy} = 37.3297$

$S_{min} = 21.9973$

df = 32

$$\hat{\sigma}^2 = 0.6874$$

$$\hat{\sigma} = 0.8291$$

Correlation coefficient: $$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0.64$$

Regression line: $$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = 0.171$$

$$\hat{\alpha} = \bar{y} - b\bar{x} = 19.45$$

Test: $\alpha = 0$

Model	**MSQ**	**df**
$H_0 : y = \beta x$	178 I527	33
$A : y = \alpha + \beta x$	21.9973	32
Reduction	156.1553	1

$$F = \frac{156.1553}{21.9973/32} = 227.16 = F(\alpha = 0)$$

$F_{.95;1,32} = 4.17$ (Table A-9)

→ Simplification not permitted

Test $\beta = 0$

Model	MSQ	df
$Ho : y = \alpha$	37.3297	33
$A : y = \alpha + \beta x$	21.9973	32
Reduction	15.3324	1

$$F = \frac{15.3324}{21.9973/32} = 22.304 = F(\beta = 0)$$

$F_{95,1,32} = 4.17$ (Table A-9)

→ Simplified model not allowed !

We obtain the following standard deviations for the parameter estimates $\hat{\alpha}$ and $\hat{\beta}$

$$sd(\hat{\alpha}) = \frac{|\hat{a}|}{\sqrt{F(\alpha=0)}} = \frac{19.45}{\sqrt{227.16}} = 1.2905$$

$$sd(\hat{\beta}) = \frac{|\hat{\beta}|}{\sqrt{F(\beta=0)}} = \frac{0.171}{\sqrt{22.30}} = 0.0362$$

and the regression line is:

$$y = \underset{(1.29)}{19.45} + \underset{(0.036)}{0.171}x \quad \hat{\sigma} = 0.8291$$

95% confidence intervals for $\hat{\alpha}$ and $\hat{\beta}$

From table A-4: $t_{.975,32} = 2.042$

$$\hat{\alpha}: 19.45 \pm 2.042 * 1.29 = (16.82, 22.08)$$

$$\hat{\beta}: 0.171 \pm 2.042 * 0.036 = (0.097, 0.245)$$

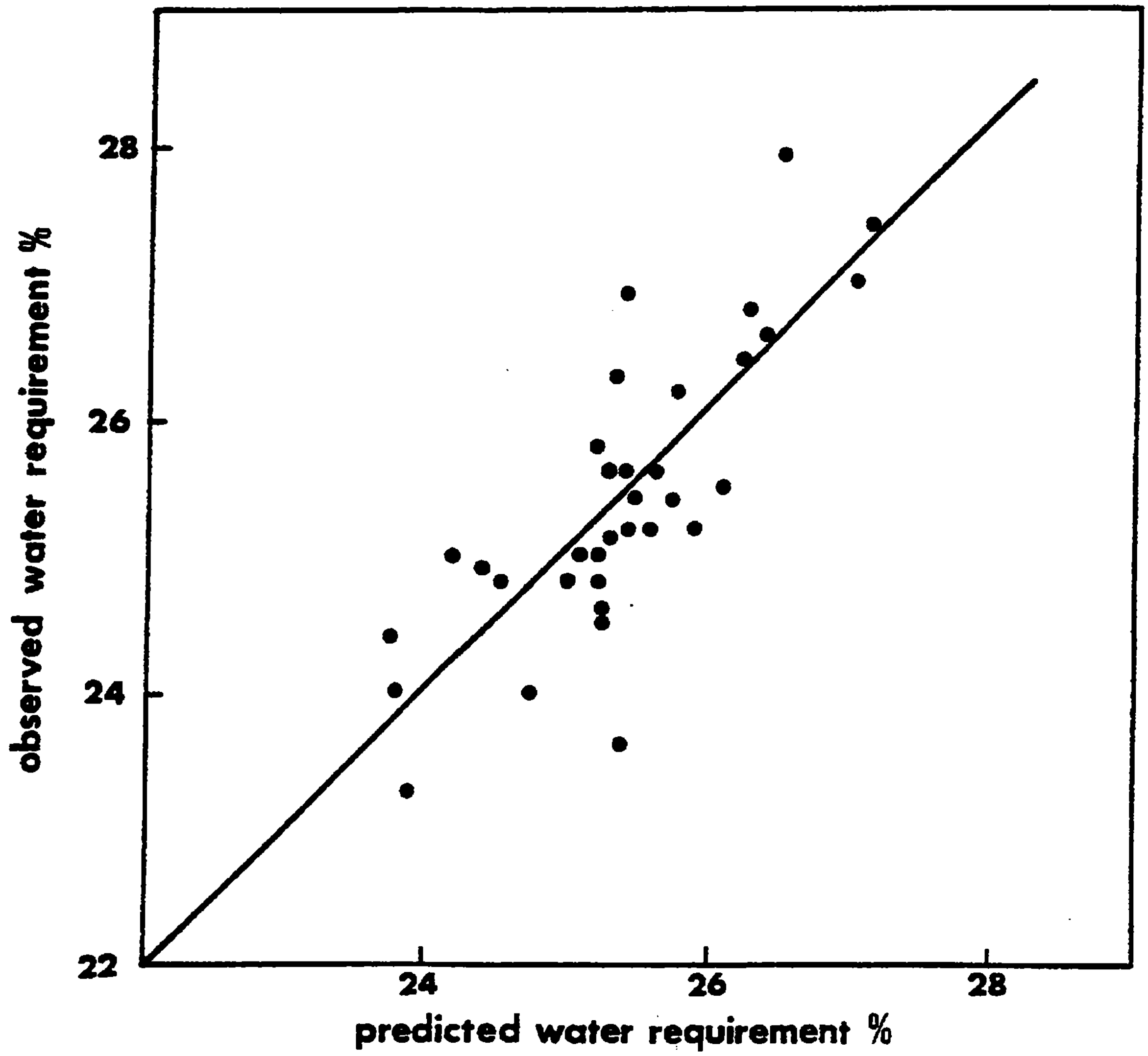
Coefficient of determination: $r^2 = 0.41$

Comment:

Only 41% of the total variation of Y can be explained by the dependence of the grain fraction. The prediction can be improved when further variables are considered in a multiple regression. The coefficient of determination increases to 69% if C_3A- and alkali-content is included in the equation

$y = 17.44 + 0.15x_1 + 0.26\, x_2 + 0.52x_3$
x_1 = grain fraction 10-32µ%
x_2 = C_3A- content %
x_3 = $K_20 + Na_2O$%

The improvement is visible in a scatter diagram of observed against predicted values.



8. STATISTICAL INVESTIGATIONS AND THEIR DESIGN

In practice numerous problems and questions cannot be answered unless special quantitative investigations are performed. This applies to every practical field, starting with the raw material exploitation and ending with the sale of the finished product. In many cases, it is not possible to attain complete information as an effective basis for decision making. For this reason, it is becoming more and more common to resort to statistical investigations in form of sample surveys and experiments.

The proper performance and interpretation of statistical investigations is essential. The so-called "lies of statistics" generally have to be attributed to the wrong selection of data, a misinterpretation, or procedures which do not relate to the objective. Experience has shown that very rarely data are manipulated intentionally. However, people who employ "wrong statistics", are convinced, due to a lack of in-depth knowledge, that their argumentation is objective and correct.

8.1 The Five Phases of a Statistical Investigation

Despite the variety of potential applications, a statistical investigation can roughly be divided into five phases in the following order:

1) Formulation of problem:
   exact definition of purpose for which information is acquired
2) Planning:
   purpose-oriented planning of investigation according to statistical principles with the objective to acquire optimum information at given expenses. In case of doubt and problems contact an experienced statistician.
3) Performance:
   procurement of data strictly in accordance with the established planning
4) Evaluation:
   summary and presentation of results; inference from sample to population
5) Conclusions:
   realistic interpretation of results

Essential basic principle:

Statistical investigations will only provide truly efficient results if knowledge in the investigated field is optimally combined with thorough statistical knowledge. Usually this will lead to teamwork, since the investigator who lacks extensive knowledge in statistics is usually just as unsuccessful in his attempt to carry out the investigation on his own as the statistician who is entrusted with the entire problem complex.

8.2 Sample Surveys and Experiments

Statistical investigations are subdivided into two groups sample surveys and experiments - which differ significantly with respect to their objective and interpretation.

Experiment

In an experiment response variables are investigated, which depend on various factors. The results are produced by a controlled variation of the factors of primary interest, whereas all other effects on the response have to be eliminated.

Factors that are being studied may be quantitative (e.g. temperature, concentration) or qualitative (e.g. type of ample preparation, origin of samples).

In an experiment, causal effects on the response variable can be determined quantitatively if the necessary provisions have been made in the planning and performance.

Sample survey

Samples are taken from one or several populations (the set of sample units is called sample of size n). Certain interesting characteristics are measured in order to evaluate (estimate)unknown characteristics of the populations or to compare populations. Contrary to the experiment, the results of a sample survey do not permit any conclusions regarding cause and effect. This is often ignored when correlations between two characteristics are assessed.

Reliable results of both experiments and sample surveys can only be obtained if the respective statistical principles have been adhered to in the planning and performance of the investigation.

8.3 Fundamental Principles in Statistical Investigations

8.3.1 Experiments

The five criteria of a good experiment are:

1) The experiment should serve a well defined purpose
In a first step the problem has to be clearly identified for each experiment, and the hypotheses to be investigated have to be determined.

2) Factors. which are not of primary interest. should not influence the results
Influencing factors, which are not included in the investigation, must be under control at fixed levels. Otherwise several effects become mixed up and cannot be separated by statistical methods. Systematic effects of factors, which cannot be controlled, are eliminated by random allocation of samples to treatments or factor levels and by random measuring sequence (randomization).
Practical performance of randomization: In example A 1 we assign a number to each sample, No. 1 - 8 to tablets and No. 9 - 16 to beads. Each number is written on a leaflet and mixed in a box. The order of measurement is given by blind drawing of the leaflets from the box.

3) The experiment should be free of systematic errors
This requirement is partially connected with 2). Systematic errors, due to any change of effects in time, are eliminated by randomization. An often underestimated source of systematic errors is the prejudice of the experimenter. In order to avoid this, mainly "blind" experiments should be performed. The samples are coded with random numbers, the decoding key only known to a confidential person, who herself is not in a position to perform analyses.

4) The experiment should provide a measure of its precision
An estimate of the precision is obtained by a replication of the experiment. For each step

of the experiment several measurements should be taken. In example A 1 it would be impossible to state whether a systematic difference exists, if only one tablet and one bead are measured. (Appendix I)

5) Precision and efficiency of an experiment should be high enough to reach the set goals
There are various measures to improve the precision and efficiency of an experiment:

- Reduction of variability by using homogeneous materials and by carefully controlling all the factors, as well as by strictly observing the analytical regulations.
- Increasing the number of replications; this will lead to the problem of determining the sample volume required to achieve the desired precision.
- Blocking: Measurements are performed in homogeneous groups (blocks). Such blocks may be 'measurements made by the same operator' or 'measurements made on the same day'.
  A special blocking procedure is the paired comparison. In example A 3 both laboratories measure the compressive strength on samples drawn from the same cement bag.
  Block experiments should be balanced and randomized, i.e. the number of measurements and treatments is equal in each block and assignment of samples or treatments to blocks is random.
  Special cases of blocking: paired comparison.
- Analysis of covariance: factors, which are not subject to the experimenter's control, are recorded in order to eliminate their effects in a later analysis.

The principles above are valid for any experimentation. Special designs for various problems are given in the literature. Problems may be: Effects of one or several factors on a response, additivity of factor effects, splitting of components of variability in a measurement procedure in order to achieve a prescribed precision at minimum cost, inter-laboratory tests, calibration, evaluation of systematic errors etc.

Note:

In contrast to a regression analysis with a given set of observations, a carefully designed experiment allows the evaluation of causal relationships. In regression the effects are usually mixed and can only be separated due to some mathematical model and not due to their real origin (ref. Interpretation of r_{xy} section 7.1).

STRUCTURE OF AN EXPERIMENT

1) Recognize problem
2) Describe problem in detail and determine hypotheses
3) Define experimental area; contact statistician; determine factors, levels and number of replications; elaborate exact analytical regulations; determine a reference basis if required. It is often necessary to analyse what happens if no treatment is applied (e.g. Placebo in clinical experiments).
4) Establish experimental design, taking into account methods to improve precision (blocking).
Randomization:
- random selection of sample units
- random allocation of treatments
- coding of samples
5) Specify variables, which cannot be maintained constant (covariables)
6) Determine number of replications
7) Determine procedural organization
8) Determine procedures of statistical data analysis; exact regulation concerning the manner in which data are to be supplied (form)
9) Perform experiment
10) Evaluate results (statistical data analysis)
11) Draw conclusions
12) Take measures

Step 2) to 7) are designated as experimental planning. The results of this planning should absolutely be recorded in an experimental plan. Since experiments are usually quite expensive, careful planning will definitely be worthwhile.

The request to prepare a written experimental plan entails significant advantages:

- the necessity to formulate statements in precise terms
- clear conditions and thus less uncertainty in the performance of the experiment
- the circumstances of the experiment can be reconstructed, if later on the results are used for comparison with new results.

8.3.2 Sample Surveys

Sample surveys and experiments are performed in an analogous way. Differences are due to the different situation. While the purpose of an experiment is to directly produce results, sampling is carried out to analyse existing characteristics of a population.

Special attention should be paid to the following points:

1) Target population and sampled population
The target population is the aggregate about which the investigator is trying to make inferences from his sample. It is usually helpful to focus the attention on differences between the population actually sampled and the population that is attempted to be studied.

2) <u>Sample unit</u>
The sample units have to be described in clear terms. They should not overlap and the sum of all the sample units must be equal to the investigated population.
In the cement production for instance, if an individual sample is analysed, the results and interpretations are different from those of an analysis performed on a daily composite sample.

3) <u>Representativity and sampling method</u>
Samples only supply unbiased results if they represent the entire population. It is often mistakenly believed that the large sample size n ensures its representativity, or it is simply maintained that the sample is representative in order to prevent rejection of the results.
Representativity can only be ensured by an appropriate, random sampling method, giving each sample unit an equal chance to be included in the sample (exceptions in special cases: sampling with unequal probabilities, systematic sampling).

In literature a number of selection procedures are described. Depending upon the problem situation, various methods ma y be considered. They differ in their practicability, financial consequences (expenses) and efficiency.

In most cases an optimum selection procedure for a specific problem situation can be found.

<u>STRUCTURE OF A SAMPLE SURVEY</u>

1) Recognize problem. What information is required about what populations?
2) Detailed definition of problem
   - definition of target populations
   - what information with what accuracy?
3) Determine the sample population. Where, when and how are the sample units extracted from the population? Determine the variables to be recorded
4) Determine sampling method (equal probability sampling, stratified sampling, cluster sampling, multi-phase sampling, multi-stage sampling, systematic sampling, unequal probability sampling)\
5) Determine sample scheme (in accordance with the required accuracy)
6) Outline procedural organization
7) Determine procedures of statistical data analysis; exact regulation concerning the manner in which data are to be supplied
8) Perform sampling
9) If possible perform measurements in random sequence
10) Evaluation and presentation
11) Conclusions
12) Take measures

9. OUTLOOK

This section will give a short outlook to further and perhaps more sophisticated statistical applications. It also contains references for further readings. All these applications require more than elementary knowledge of statistical methods.

9.1 Time series and growth curves analysis

This is the analysis of chronological observations, (e.g. daily temperatures at a certain location or observations over a certain time period of a patient obtaining a certain medicament). The characteristic of these data are that the observations are not independent from each other and the method of linear regression cannot be applied.

Literature:

Box, G.E.P. and Jenkins, G.M. (1976). Time Series Analysis, Forecasting and Control, second edition. San Francisco: Holden-Day, Inc.

Cox, D.R. and Lewis, P.A.W. (1966). The Statistical Analysis of Series of Events. London: Methuen.

Nelson, C.R. (1973). Applied Time Series Analysis for Managerial Forecasting. San Francisco: Holden-Day, Inc.

9.2 Categorical and Qualitative Data Analysis

This is the analysis of counts. Many experiments contain qualitative and nonmetric variables, e.g. when analyzing the number of smokers and non-smokers beyond male and female persons. Smoking behaviour and sex have no natural numeric values associated with them. Used statistical methods for such data are

- Crosstabulation
- Contingency tables
- Goodness of fit tests
- Log linear models
- Logistic regression
- Correspondence analysis

Literature:

Agresti, A. (1984). Analysis of ordinal categorical data. New York: Wiley-Interscience.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge, MA: MIT Press.

Haberman, S.J. (1978). Analysis of Qualitative Data, Vol. 1: Introductory Topics. New York: Academic Press.

Haberman, S.J. (1979). Analysis of Qualitative Data, Vol. 2: New Developments. New York: Academic Press.

9.3 Experimental Designs and ANOVA

This is a wide field containing among others:

- One-Way ANOVA
- Multifactor ANOVA
- Analysis of Nested Designs
- Full and Fractional Designs
- Response Surface Analysis
- Covariance Analysis
- and a lot more.

Literature:

Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978). Statistics for Experimenters. New York: Wiley.

Neter, J. and Wassermann, W. (1974). Applied Linear Statistical Models. Homewood, Illinois: Richard E. Irvin, Inc.

9.4 Multivariate Methods

This title involves the analysis of multivariate data. It is not appropriate to analyze multivariate data univariate, because correlations among the different variables are not taken into account.

Statistical Methods:

- Correlation Analysis
- Covariance Analysis
- Multiple Regression
- Principal Components
- Factor Analysis
- Discriminant Analysis
- Cluster Analysis
- Canonical Correlations
- Multidimensional Scaling
- and a lot more.

Literature:

Sieber, G.A.F. (1984). Multivariate Observations. New York: Wiley.

Flury, B. and Riedwyl, H. (1988). Multivariate Statistics. A practical approach. London, New York: Chapman and Hall.

Johnson, R.A. and Wichern, D.W. (1982). Applied Multivariate Statistical Analysis. London: Prentice-Hall.

Morrison, D.F. (1976, 2nd edition). Multivariate Statistical Methods. New York: McGraw-Hill.

Everitt, B.S. (1980). Cluster Analysis, 2nd edition, London: Heinemann Education Books, Ltd.

9.5 Nonparametric Methods

Parametric methods base on certain assumptions on the data (e.g. normality of residuals in linear regression, normality of the observations in testing etc.). If these assumptions do not hold, it is often more efficient to use methods that do not use a particular underlying distribution function, so called nonparametric methods (e.g. the Wilcoxon Test known for Section 5.2.4 is a nonparametric test). Some nonparametric procedures exist for:

- Tests of binary sequences
- Tests for randomness
- Tests for location
- Comparison of 2 samples
- Rank correlation analysis
- Goodness of fit tests

Literature:

Conover, W.J. (1980). Practical nonparametric statistics, 2nd edition. New York: John Wiley and Sons, Inc.

Lehmann, E.L. (1975). Nonparametrics. San Francisco: Holden-Day, Inc.

Gibbons, J.D. (1976). Nonparametric Methods for Quantitative Analysis. New York: Holt, Rinehart and Winston.

Hollander, M. and Wolfe, D.A. (1973). Nonparametric Statistical Methods. New York: Wiley.

9.6 Bootstrap and Jack-knife Methods

In dealing with complicated functions of data, it is mostly not possible to derive the underlying distribution function. Bootstrap and Jackknife sometimes provide the possibility to derive the distribution functions, statistical measures and a lot more at such complicated data.

Literature:

Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. Philadelphia: Soc. for Industrial and Applied Math.

9.7 Simulation and Monte Carlo Method

By use of simulation and Monte Carlo methods it is often possible to investigate and analyse complex functions of random variables or systems of events. Simulation incorporates also the generation of random variables.

Literature:

Fishman, G.S. (1978). Principles of Discrete Event Simulation. New York: John Wiley & Sons, Inc.

Rubinstein, R.Y. (1981). Simulation and the Monte Carlo Method. New York: John Wiley & Sons, Inc.

9.8 General Literature

Seber, G.A.F. (1984). Multivariate Observations. New York: Wiley.

Belsley, D.A. Kuh, E. and Welsh, R.E. (1980). Regression Diagnostic: Identifying Influential Data and Sources of Collinearity. New York: Wiley.

Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P.A. (1983). Graphical Methods for Data Analysis. Boston: Duxbury Press.

Flury, B. (1980). Construction of an asymmetrical face to represent multivariate data graphically. Tech. Rep. No. 3, University of Berne, Dept of Statistics.

Schupbach, M. (1984). ASYMFACE - Asymmetrical Faces on IBM-PC. Tech.Rep. No.16. University of Berne, Dept of Statistics.

Draper, N.R. and Smith, H. (1981), 2nd ed.). Applied Regression Analysis. New York: Wiley.

Grant, E.L. and Leavenworth, R.S. (1980). Statistical Quality Control, fifth ed. New York: McGraw Hill.

Montgomery, D.C. (1985). Introduction to Statistical Quality Control. New York: Wiley.

Recommended literature pp. 90 ff

Huff, D. (1974). How to lie with statistics.

Miller, R.G. (1981). Simultaneous Statistical Inference, 2nd ed. New York: Wiley.

Morrison, D.F. (1983). Applied Linear Models. Englewood Cliffs, NJ: PrenticeHall Inc.

10. STATISTICAL PROGRAM PACKAGES

Very complete and sophisticated packages are (all programs are available for PC's under DOS):

BMDP	Statistical Software Ltd., Cork Technology Park. Cork, Ireland (phone: 021-542722)
SAS:	SAS Inst. GmbH, Cary NC, USA (phone: (919)467-8000)
SPSS	SPSS Inc., Chicago Il, USA (phone: (312)329-3300)
SYSTAT	SYSTAT Inc., Evanston Il, USA (phone: (312)864-5670)
STATGRAPHICS	STSC Inc., Rockville MD, USA (phone: (301)984-5000)

A review of 49 statistical packages is given in the PC magazine, Vol. 8, No. 5, March 1989.

Appendix I

Example A1:

In the application of XRF-analysis of raw meal it was required to know to what extent the results are dependent on sample preparation, by pressed powder tablets and by fusion with Litetraborate, respectively.

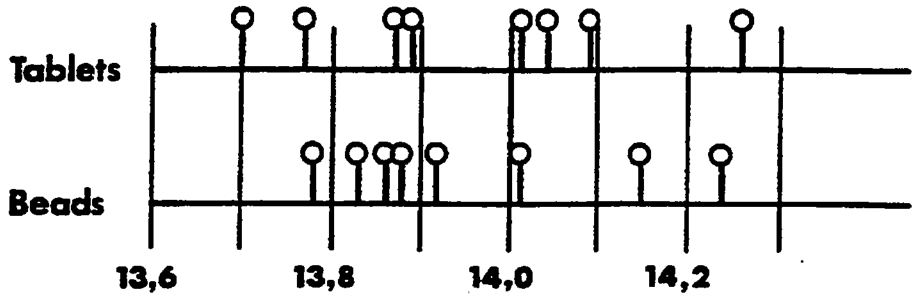
With 8 preparations of tablets and beads of the same raw meal the following results for CaO and SiO_2 were obtained.

CaO	
Tablets	Beads
42.63	42.42
42.59	42.44
42.63	42.45
42.80	42.42
43.02	42.59
42.61	42.49
42.40	42.62
42.79	42.26

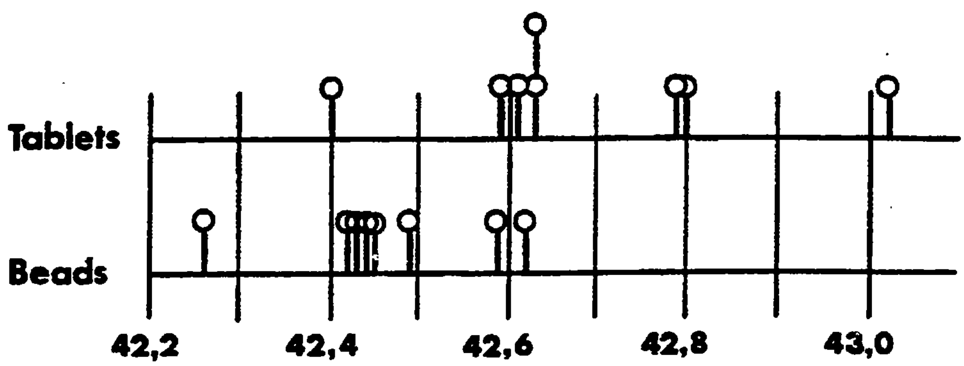
SiO2	
tablets	beads
13.87	14.15
14.01	13.86
14.04	13.92
14.09	13.83
14.26	14.01
13.70	13.78
13.89	14.24
13.77	13.88

We are interested in differences between the preparation methods. To get a quick survey on the data we mark every observation on the measurement scale for tablets and beads respectively.

SiO_2



CaO



Looking at the graph we suppose a significant difference of CaO-results between beads and tablets. For Si 02 results no difference is visible.

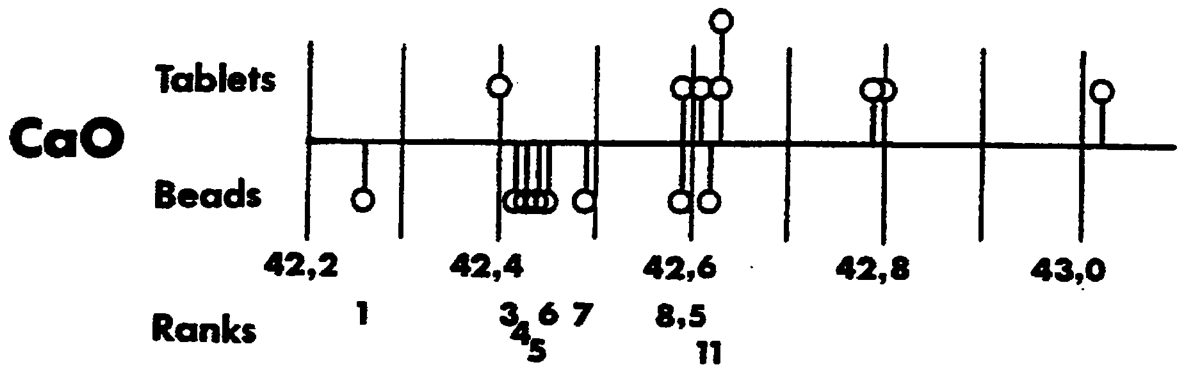
The final decision whether the differences are significant is made with an appropriate test.

Test situation:

- We test the hypothesis that the mean results of both preparation methods are the same H_0: $\mu_{tablets} = \mu_{beads}$
- Two independent samples with small sample sizes N = 8
- two-sided test: we have no idea in what direction a possible difference may occur before the observations are taken.

Test procedure: Wilcoxon-test at α = 5% level

As an alternative representation we can mark the data on both side of one line. In this case ranks of the observations can be read directly from the graph.



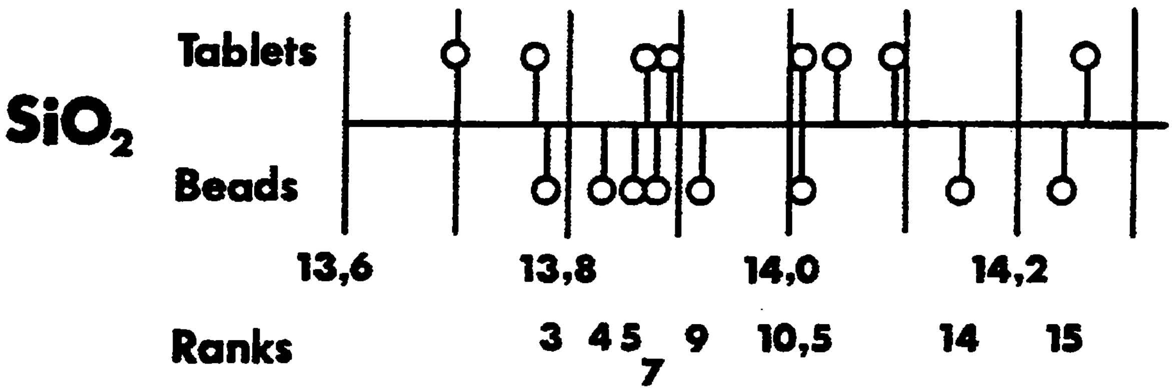
The ranksum of beads is R = 45.5

$$W = 2 . 45.5. - 8 . 17 = -45$$

$$|W| = 45 => 38 = W_{0.975}$$ (from Table A-8, Appendix III).

Decision:

The difference of CaO results between tablets and beads is significant.



$$|W| = 1 < 38 = W_{0.975}$$

There is no significant difference for SiO_2 results.

Example A2:

Control of homogenization efficiency.

In intervals of 30 minutes, 20 samples are taken from the raw ma serial stream both before and after homogenization.

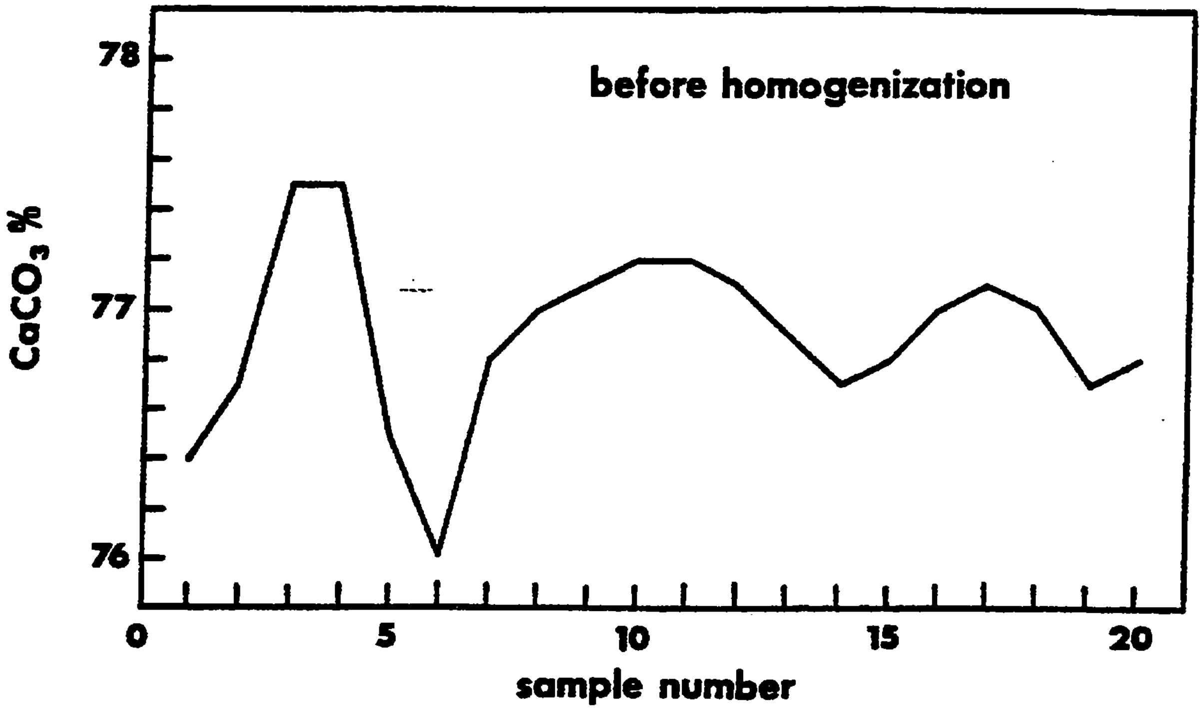
Measured values: $CaCO_3$ content

Total capacity of homogenization silo: 1000 tons

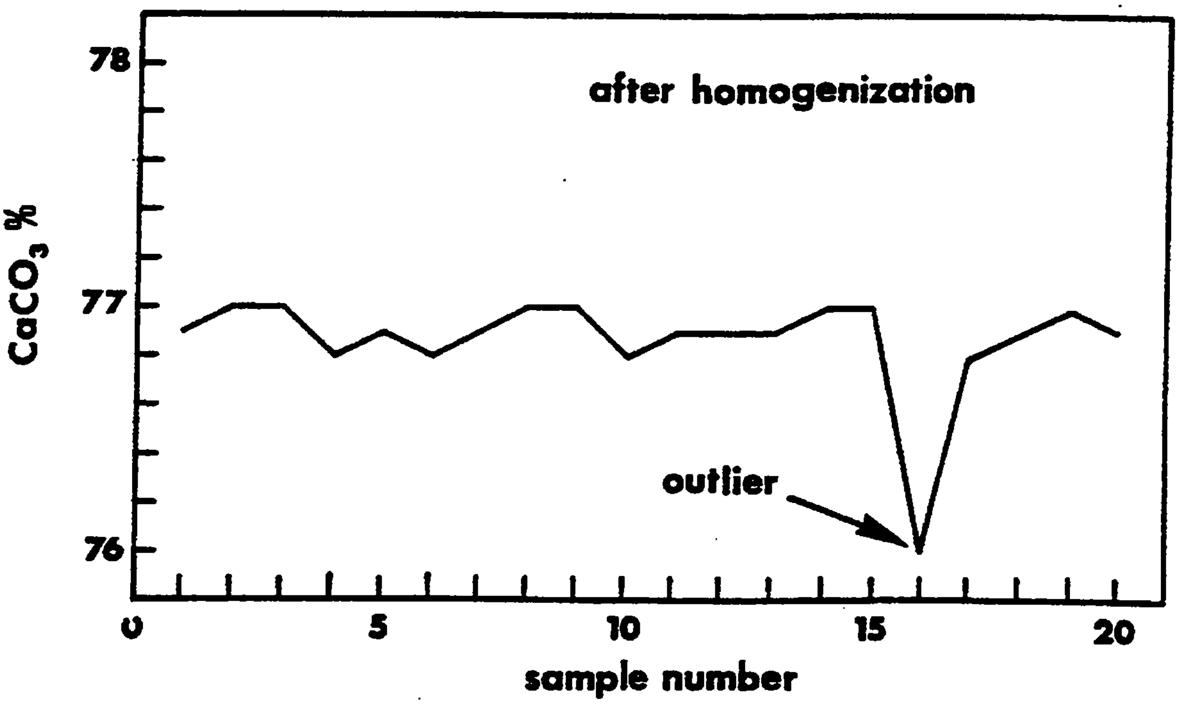
Sample No.	$CaCO_3$ before homogenization	$CaCO_3$ after homogenization
1	76.4	76.9
2	76.7	77.0
3	77.5	77.0
4	77.5	76.8
5	76.5	76.9
6	76.0	76.8
7	76.8	76.9
8	77.0	77.0
9	77.1	77.0
10	77.2	76.8
11	77.2	76.9
12	77.1	76.9
13	76.9	76.9
14	76.7	77.0
15	76.8	77.0
16	77.0	76.0
17	77.1	76.8
18	77.0	76.9
19	76.7	77.0
20	76.8	76.9

The graph is chronological order (time-plot) shows a systematical variation of $CaCO_3$ content before homogenization. Except for one outlying value the variation is small after homogenization.

Before Homogenization



After Homogenization



Using the Dixon criterion to check for outliers we compute

$$r_{22} = \frac{x_{(3)} - x_{(1)}}{x_{(n-2)} - x_{(1)}} = \frac{76.8 - 76.0}{77.0 - 76.0} = 0.8$$

with $x_{(1)}$ the smallest, $X_{(3)}$ the third smallest and $X_{(n-2)}$ the largest value.

Since $r_{22} = 0.8 > r_{0.995} = 0.562$ (Table A-2, Appendix III) the extreme value is considered to be a real outliner (with significance level $\alpha = 1\%$). It is deleted for further analysis.

Statistical data description

	$CaCO_3$ before homogenization	**$CaCO_3$ after homogenization**	
n	20	(20)	19
$\bar{x}$	76.9	(76.87)	76.92
$\bar{x}$	76.95	(76.9)	76.9
s	0.36	(0.22)	0.076 one outlier deleted: 76.0
v	0.0046	(0.0028)	0.0010
x_{max}	77.5	(77.0)	77.0
x_{min}	76.0	(76.0)	76.8
R	1.5	(1.0)	0.2

Comment:

Mean and median are almost equal. The distribution of observations seems to be therefore symmetrical about the mean either before and after homogenization. Due to homogenization the standard deviation is reduced from 0.36 to 0.08 corresponding to a factor of four to five.

Note:

Elimination of the outliner reduces the standard deviation from 0.22. to 0.08 !

Example A3:

In order to control mortar strength a national control laboratory takes every month a sample of cement in a plant. It was supposed that these control results tend to be lower than results of internal quality control.

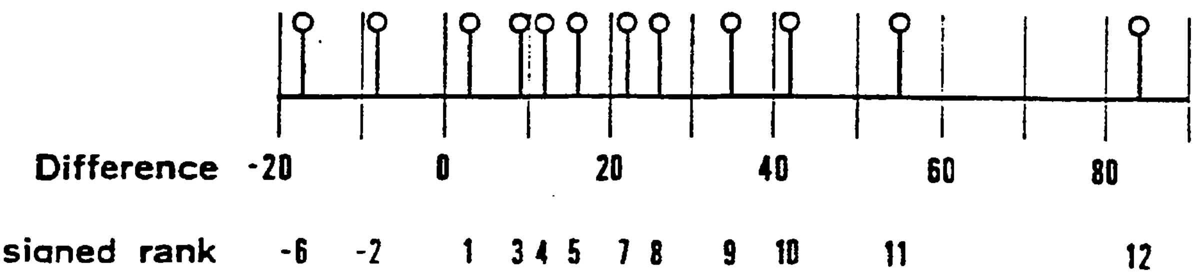
To find out a suspected syste ma tic error between results of control office and plant laboratory during a year the samples were measured in both laboratories:

	Plant x	**Control y**	**Difference d = x - y**
January	548	526	22
February	540	524	16
March	574	591	-17
Apri1	469	477	-8
May	540	531	9
June	515	431	84
July	520	485	35
August	531	476	55
September	530	490	42
October	464	452	12
November	524	498	26
December	519	516	3

The strength values are paired. Each pair of observation is concerned with the same sample. To test for a significant difference we use the singed-rank test at a significance level of $\alpha = 5\%$ (two-sided)

$$H_0: \mu = 0$$

Plot of differences $d_i = x_i - y_i$



Sum of signed ranks:

$$T = (-6) + (-2) + 1 + 3 + 4 + 5 + 7 + 8 + 9 + 10 + 11 + 12 = 62$$

Since $T = 62 > 52 = T_{0\ 975}$ we decide that the difference between results is systematic.

Example A4:

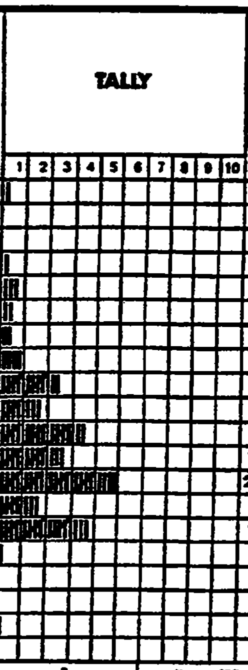
The following data represent 115 measurements of Schmidt-haamer-strength (abutment of a bridge).

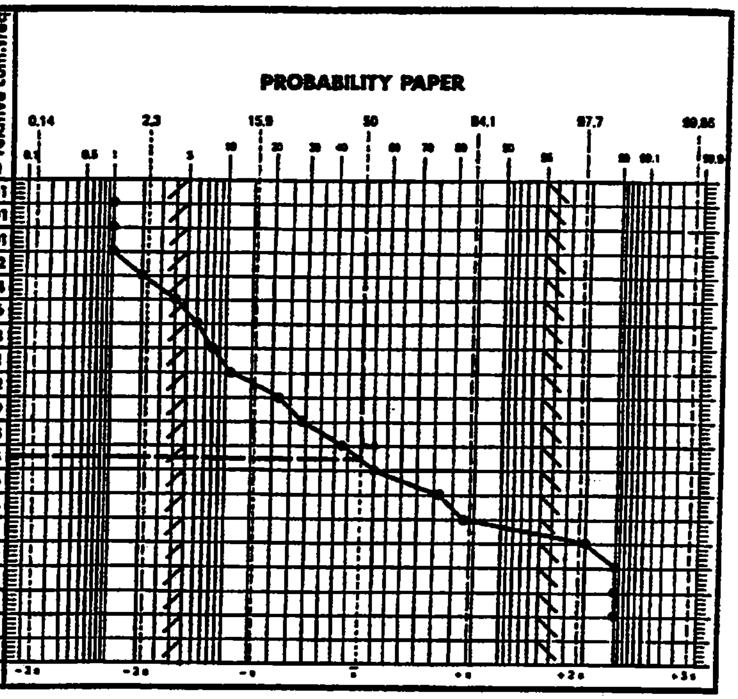
44	37	46	45	46	45	46	46	46	48	46	46	48	44
42	45	45	46	48	47	45	46	41	46	43	42	44	45
40	45	42	44	38	39	39	45	44	44	46	48	47	48
48	46	48	42	48	45	45	48	52	43	38	43	42	44
46	44	44	42	34	42	42	42	41	46	45	48	46	46
48	47	46	44	45	40	47	41	42	47	43	45	43	43
44	44	41	44	48	46	47	48	48	46	44	48	44	49
46	48	44	43	43	46	47	46	48	46	46	47	38	42
48	44	42											

A detailed interpretation of these data requires their representation in a frequency table and histogram. With the probability paper we check the data for normality.

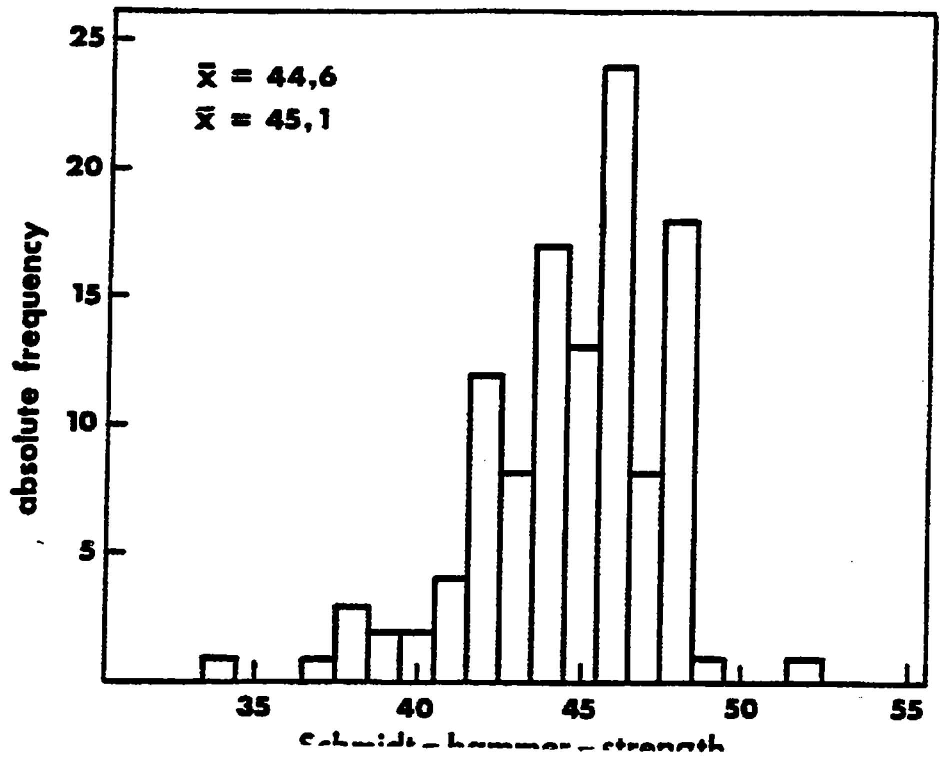
The cumulative frequency curve plotted on the probability paper is not linear and the distribution is therefore not normal.

class limits	mid-point	n_i absolute freq.	G_i cumulated freq.	H_i relative cum. freq.
	34	1	1	.01
34.5	35	0	1	.01
35.5	36	0	1	.01
36.5	37	1	2	.02
37.5	38	3	5	.04
38.5	39	2	7	.06
39.5	40	2	9	.08
40.5	41	4	13	11
41.5	42	12	25	.22
42.5	43	8	33	.29
43.5	44	17	50	.43
44.5	45	13	63	.55
45.5	46	24	87	.76
46.5	47	8	95	.83
47.5	48	18	113	.98
48.5	49	1	114	.99
49.5	50	0	114	99
50.5	51	0	114	.99
51.5	52	1	115	1.0
class width w = 1		$n = \Sigma n_i = 115$		





Histogram and tally show that even values of Schmidt-hammer strength are more frequent than odd values. This seems to be caused by a reading error of scale. Probably only even values are marked on the measurement scale.



EXTENDED PROBABILITY PAPER

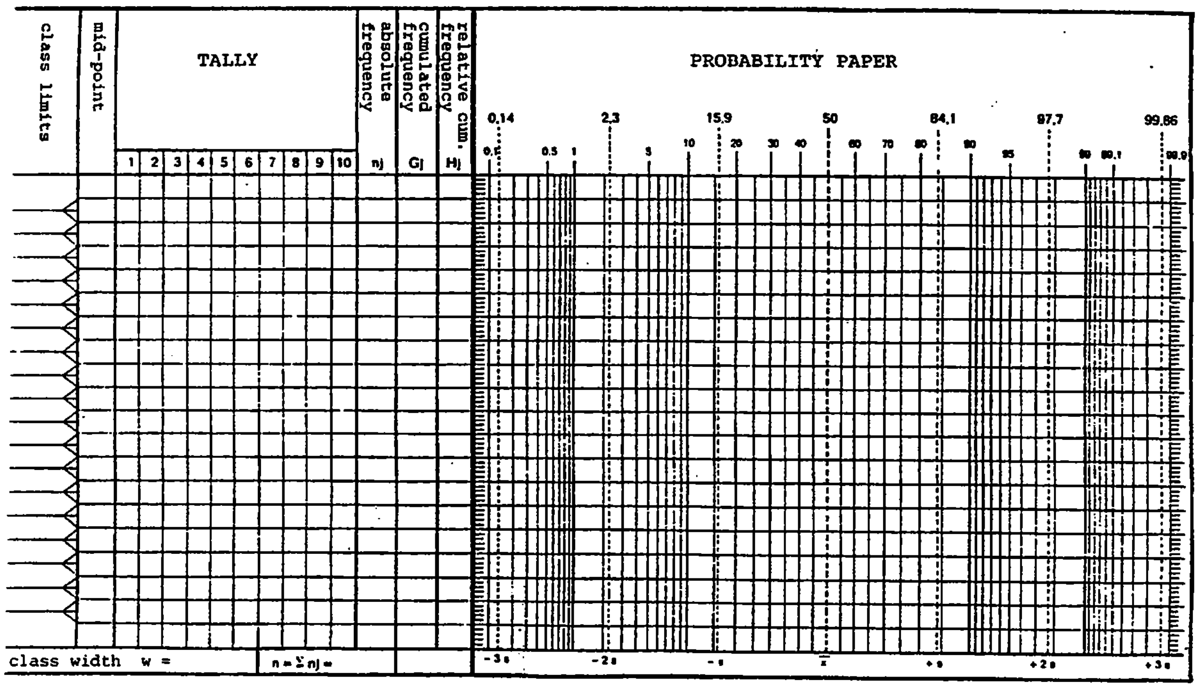
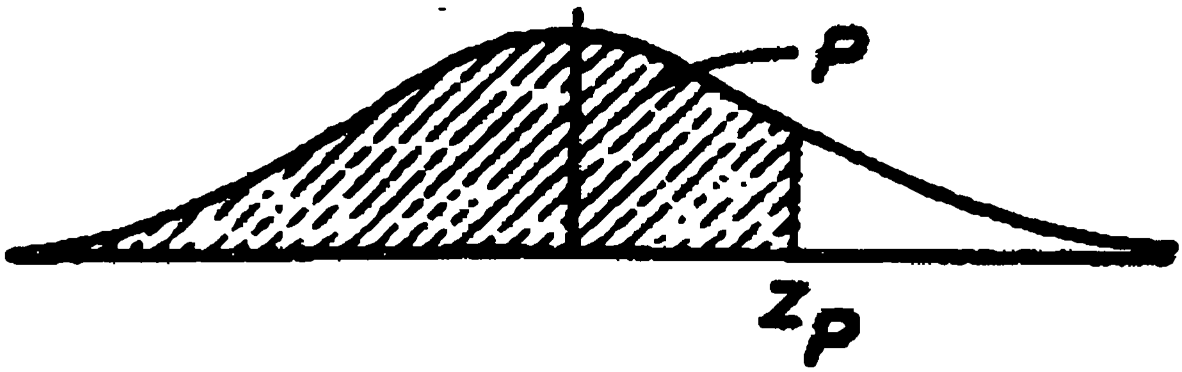


Table A-1 Standard Normal Distribution - Values of P



Values of P corresponding to z_p for the normal curve.

z is the standard normal variable. The value of P $-z_p$ equals one minus the value of P for $+z_p$, e.g. the P for -1.62 equals 1-0.9474 = .0526.

z_p	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Special Values

P	.001	.005	.010	.025	.050	.100
z_p	−3.090	−2.576	−2.326	−1.960	−1.645	−1.282

P	.999	.995	.990	.975	.950	.900
z_p	3.090	2.576	2.326	1.960	1.645	1.282

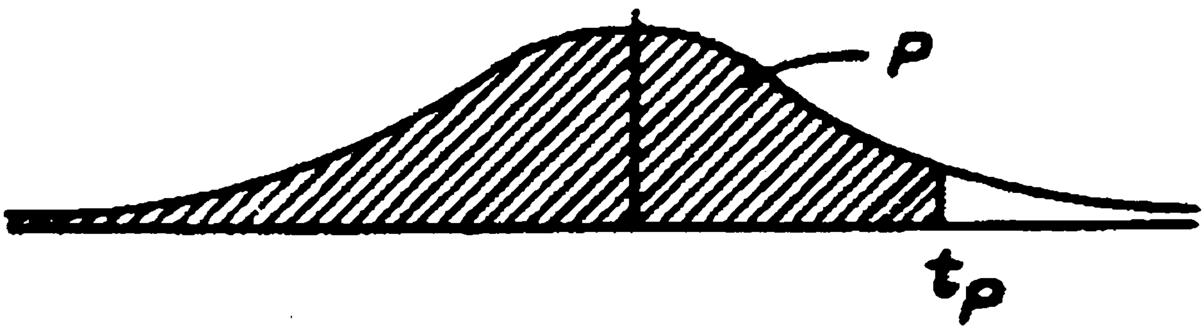
Table A-2 Critical values for the Dixon Criterion

Statistic	Number of Observations, n	Upper Percentiles						
		.70	.80	.90	.95	.98	.99	.995
r_{10}	3	.684	.781	.886	.941	.976	.988	.994
	4	.471	.560	.679	.765	.846	.889	.926
	5	.373	.451	.557	.642	.729	.780	.821
	6	.318	.386	.482	.560	.644	.698	.740
	7	.281	.344	.434	.507	.586	.637	.680
r_{11}	8	.318	.385	.479	.554	.631	.683	.725
	9	.288	.352	.441	.512	.587	.635	.677
	10	.265	.325	.409	.477	.551	.597	.639
r_{21}	11	.391	.442	.517	.576	.638	.679	.713
	12	.370	.419	.490	.546	.605	.642	.675
	13	.351	.399	.467	.521	.578	.615	.649
r_{22}	14	.370	.421	.492	.546	.602	.641	.674
	15	.353	.402	.472	.525	.579	.616	.647
	16	.338	.386	.454	.507	.559	.595	.624
	17	.325	.373	.438	.490	.542	.577	.605
	18	.314	.361	.424	.475	.527	.561	.589
	19	.304	.350	.412	.462	.514	.547	.575
	20	.295	.340	.401	.450	.502	.535	.562
	21	.287	.331	.391	.440	.491	.524	.551
	22	.280	.323	.382	.430	.481	.514	.541
	23	.274	.316	.374	.421	.472	.505	.532
	24	.268	.310	.367	.413	.464	.497	.524
	25	.262	.304	.360	.406	.457	.489	.516

Table A-3 Critical values for Outlier-tests, (large sample size n)

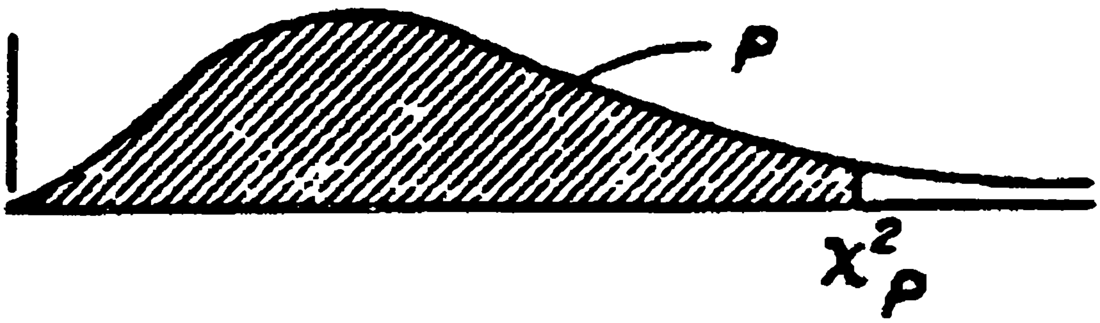
n	probability 1-α										n
	0.5%	1.0%	2,5%	5,0%	10,0%	90%	95%	97.5%	99%	99,5%	
20	2,95	3,01	3,10	3,18	3,29	4.32	4,49	4,63	4.79	4.91	20
30	3,22	3,27	3.37	3,46	3,58	4,70	4.89	5,06	5,25	5,39	30
40	3,41	3,46	3,57	3,66	3,79	4.96	5,15	5,34	5,54	5,69	40
50	3,57	3,61	3,72	3,82	3,94	5,15	5,35	5,54	5,77	5,91	50
60	3,69	3,74	3,85	3,95	4,07	5,29	5,50	5,70	5.93	6.09	60
80	3,88	3,93	4,05	4,15	4,27	5,51	5,73	5,93	6.18	6,35	80
100	4,02	4,09	4,20	4,31	4,44	5,68	5,90	6.11	6,36	6.54	100
150	4,30	4,36	4,47	4,59	4.72	5,96	6,18	6,39	6.64	6,84	150
200	4,50	4,56	4,67	4,78	4,90	6,15	6.38	6,59	6,85	7,03	200
500	5,06	5,13	5,25	5,37	5.49	6.72	6,94	7,15	7.42	7,60	500
1000	5,50	5,57	5,68	5,79	5,92	7,11	7,33	7,54	7,80	7,99	1000

Table A-4 Critical values of the t-distribution



df	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

Table A-5 Critical values of the x^2-distribution



Values of χ^2_p corresponding to P

df	$\chi^2_{.005}$	$\chi^2_{.01}$	$\chi^2_{.025}$	$\chi^2_{.05}$	$\chi^2_{.10}$	$\chi^2_{.90}$	$\chi^2_{.95}$	$\chi^2_{.975}$	$\chi^2_{.99}$	$\chi^2_{.995}$
1	.000039	.00016	.00098	.0039	.0158	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.1026	.2107	4.61	5.99	7.38	9.21	10.60
3	.0717	.115	.216	.352	.584	6.25	7.81	9.35	11.34	12.84
4	.207	.297	.484	.711	1.064	7.78	9.49	11.14	13.28	14.86
5	.412	.554	.831	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	.676	.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
120	83.85	86.92	91.58	95.70	100.62	140.23	146.57	152.21	158.95	163.64

For large degrees of freedom,

$\chi^2_P = \frac{1}{2}(z_P + \sqrt{2m - 1})^2$ approximately,

where m = degrees of freedom and z_P is given in Table A-1

Table A-6 Values of $\lambda_{1-\alpha}$ for confidence limits (Range method)

n \ α	5%	2,5%	1%	0,5%	0,1%	0,05%
2	3.157	6,353	15.910	31,828	159,16	318,31
3	0,885	1,304	2,111	3,008	6.77	9,58
4	.529	0,717	1,023	1,316	2,29	2,85
5	,388	,507	0,685	0,843	1,32	1,58
6	0,312	0,399	0,523	0,628	0,92	1,07
7	,263	,333	.429	,507	.71	0,82
8	,230	,288	.366	.429	,59	,67
9	,205	,255	,322	,374	,50	,57
10	,186	,230	,288	,333	,44	,50
11	0,170	0,210	0,262	0,302	0.40	0.44
12	,158	,194	,241	,277	,36	,40
13	,147	,181	,224	,256	,33	,37
14	,138	,170	,209	,239	,31	,34
15	,131	,160	,197	,224	,29	,32
16	0,124	0,151	0,186	0,212	0,27	0,30
17	,118	,144	,177	,201	,26	,28
18	,113	,137	,168	,191	,24	,26
19	.108	,131	,161	.182	,23	,25
20	,104	,126	,154	,175	,22	,24

[1] Nach E. Lord. The use of range in place of standard deviation in the *t*-test. Biometrika 34, 1947, S. 41.

Table A-7 Critical values of the signed-rank test

number of differences not equal zero: m	$T_{0,95}$	$T_{0,975}$
5	7	-
6	8	10
7	11	12
8	13	15
9	14	17
10	17	19
11	20	23
12	22	26
13	24	28
14	27	31
15	30	35
16	33	39
17	35	42
18	38	45
19	42	49
20	45	53
21	48	57
22	51	60
23	55	65
24	59	69
25	62	73

Table A-8 Critical values of the Wilcoxon-test

Values of $W_{0.95}$

n_2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	size of smaller sample n_1																		
20	32	38	44	50	56	62	66	72	76	82	86	92	96	100	106	110	114	120	124
19	30	37	42	49	54	59	64	69	74	79	84	87	92	97	102	105	110	115	
18	28	36	40	46	52	56	62	66	70	76	80	84	88	94	98	102	106		
17	28	33	38	45	50	53	58	63	66	73	76	81	84	89	94	97			
16	26	32	36	42	46	52	56	60	64	68	72	78	82	86	90				
15	24	31	36	39	44	49	54	57	62	65	70	73	78	81					
14	22	28	34	38	42	46	50	54	58	62	66	70	74						
13	22	27	32	35	40	43	48	51	56	59	62	67							
12	20	26	30	34	38	42	44	48	52	56	60								
11	20	23	28	31	34	39	42	45	48	53									
10	18	22	26	28	32	36	40	42	46										
9	16	21	24	27	30	33	36	39											
8	14	18	22	24	28	30	34												
7	14	17	20	23	26	27													
6	12	14	18	20	22														
5	10	13	16	17															
4	—	12	14																
3	—	9																	

Values of $W_{0.975}$

n_2	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	size of smaller sample n_1																		
20	36	44	52	60	66	72	78	84	90	96	102	108	114	120	124	130	136	142	146
19	34	43	50	57	64	69	76	81	86	93	98	103	110	115	120	125	130	135	
18	32	40	48	54	60	66	72	78	84	88	94	100	104	110	116	120	126		
17	30	39	46	51	58	63	68	75	80	85	90	95	100	105	110	115			
16	30	36	42	50	54	60	66	70	76	82	86	90	96	100	106				
15	28	35	40	47	52	57	62	67	72	77	82	87	92	97					
14	26	32	38	44	50	54	60	64	68	74	78	82	86						
13	24	31	36	41	46	51	56	61	64	69	74	79							
12	22	28	34	38	44	48	52	56	62	66	70								
11	22	27	32	37	40	45	50	53	58	61									
10	20	24	30	34	38	42	46	50	54										
9	18	23	28	31	34	39	42	47											
8	16	20	24	28	32	36	38												
7	—	19	22	25	30	33													
6	—	16	20	24	26														
5	—	15	18	21															
4	—	—	16																
3	—	—																	

Table A-9 (continued) Critical values of the F-distribution : $F_{0.975}$

$F_{.975}\ (m_1,\ m_2)$

m_1 = degrees of freedom for numerator; m_2 = degrees of freedom for denominator

$m_2 \backslash m_1$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1:67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

Recommended Literature

The following list gives a selection of books on applied statistics for further reading. The list is restricted to books which are easily comprehensible for users without profound knowledge in mathematics.

C. Chatfield — "Statistics for Technology" Chapman and Hall, London (1975), 350 pages
General survey on statistical methods with good comments on the interpretation of statistical results. Special chapters on regression, design of experiments and quality control

Noether — "Introduction to Statistics: A Fresh Approach" Houghton Mifflin Company, Boston (1971) 230 pages
Modern statistical methods, especially for the analysis of experiments. Good explanation of basic statistical ideas and problems based on intuition.

M.G. Natrella — "Experimental Statistics" US Department of Commerce, NBS Handbook 91 (1963)
"Cookbook" with many test procedures and designs of experiments. Often it is somewhat difficult to know what procedure has to be chosen.

M.R. Spiegel — "Theory and Problems of Statistics" Schaum's Outline Series,(1961) New York 350 pages
Definitions and procedures of classical statistics accompanied by many examples (875 solved problems)